Introduction to Data and Data Science

Engr. Elisa G. Eleazar

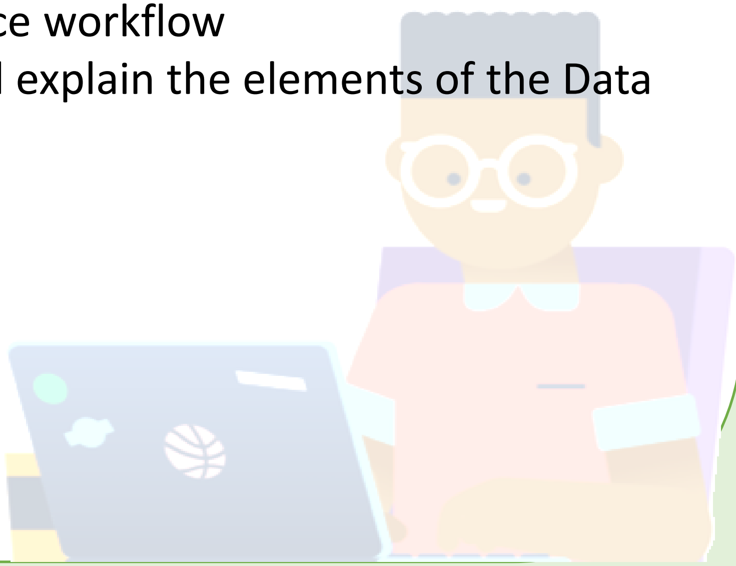
School of Chemical, Biological, and Materials Engineering and Sciences

Outline

Module 1.1: INTRODUCTION TO DATA AND DATA SCIENCE
Data and the Data Ecosystem
Data Science and its Applications
Data Science Roles and Tasks
The Data Scientist
The Data Science Workflow
Data Collection and Storage
Preparation, Exploration and Visualization
Experimentation and Prediction

Learning Outcomes

1. Define data, its properties, importance and capabilities
2. Explain the drivers of data and the current data ecosystem
3. Define Data Science and differentiate its applications
4. Differentiate the Data Science roles and enumerate the tools needed by each role
5. Explain the skills and characeristics that a Data Scientist must possess
6. Explain the phases of the Analytics lifecycle and relate these to the Data Science workflow
7. Identify, enumerate and explain the elements of the Data Science workflow



Data and the Data Ecosystem

- information collected for use
- information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer

- actual information (such as measurements or statistics) used as a basis for reasoning, discussion or calculation
- information in digital form that can be transmitted or processed
- information output by a sensing device or organ that includes both useful and irrelevant information and must be processed to be meaningful

Merriam-Webster

- information in raw or unorganized form (such as alphabets, numbers, or symbols) that refer to, or represent, conditions, ideas or objects
- symbols or signals that are input, stored, and processed by a computer, for output as usable information

- a set of values of subjects with respect to qualitative or quantitative variables

Data becomes information when it is viewed in context or in post-analysis.

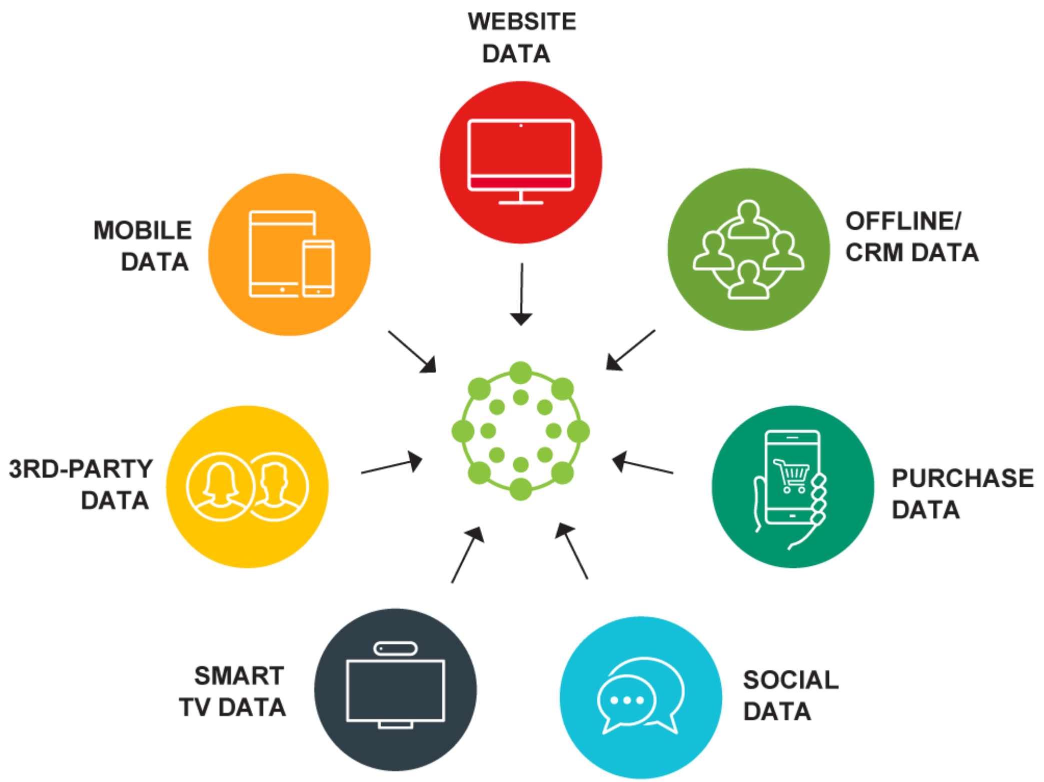
Data and the Data Ecosystem

WHAT CAN DATA DO?

- describe the current state of an organization or process (i.e., energy consumption)
- detect anomalous events (i.e, fraudulent purchases)
- diagnose the causes of events and behaviors (i.e., Spotify or Netflix activity)
- predict future events (i.e, forecasting population size)

Data and the Data Ecosystem

WHAT CAN DATA DO?



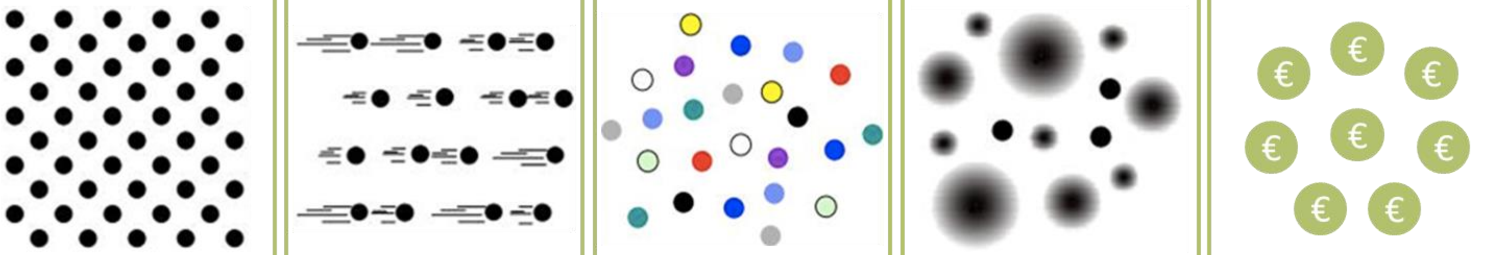
- What data should be collected?
- What methods are there for reasoning from data?
- How do we get answers from the data to answer our most pressing questions about our businesses, our lives and our world?

Data and the Data Ecosystem

BIG DATA

Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value. – McKinsey Global Report (2011)

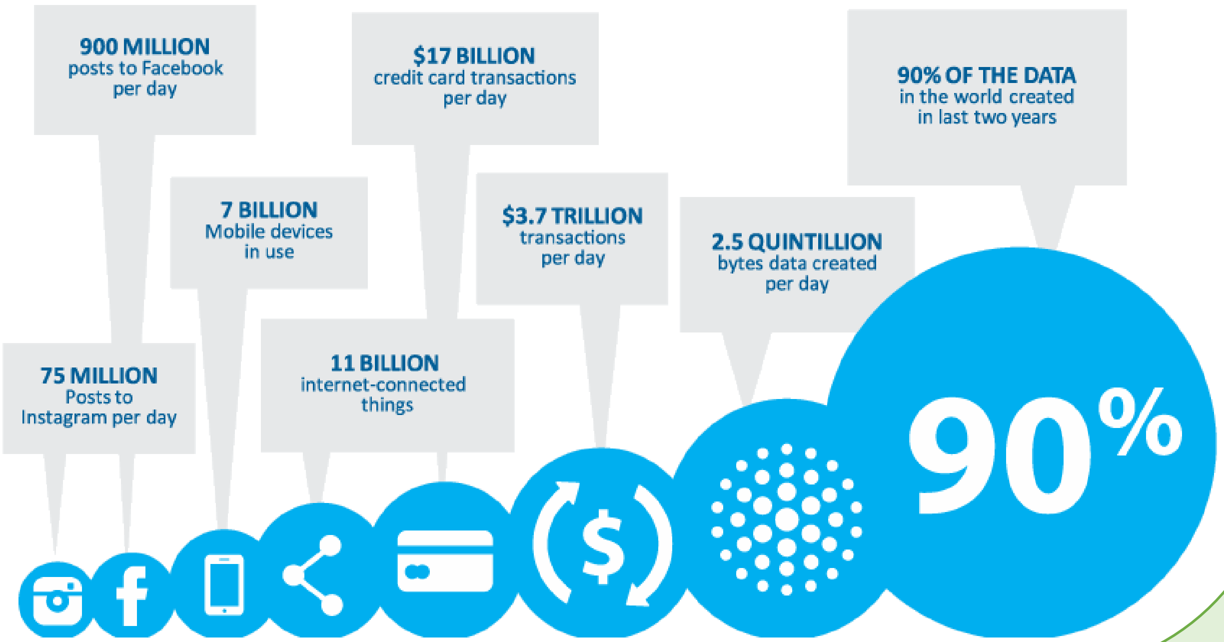
Volume	Velocity	Variety	Veracity	Value
Data at Rest	Data in Motion	Data in Many Forms	Data in Doubt	Data into Money
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,…	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data



Adapted by a post of Michael Walker on 28 November 2012

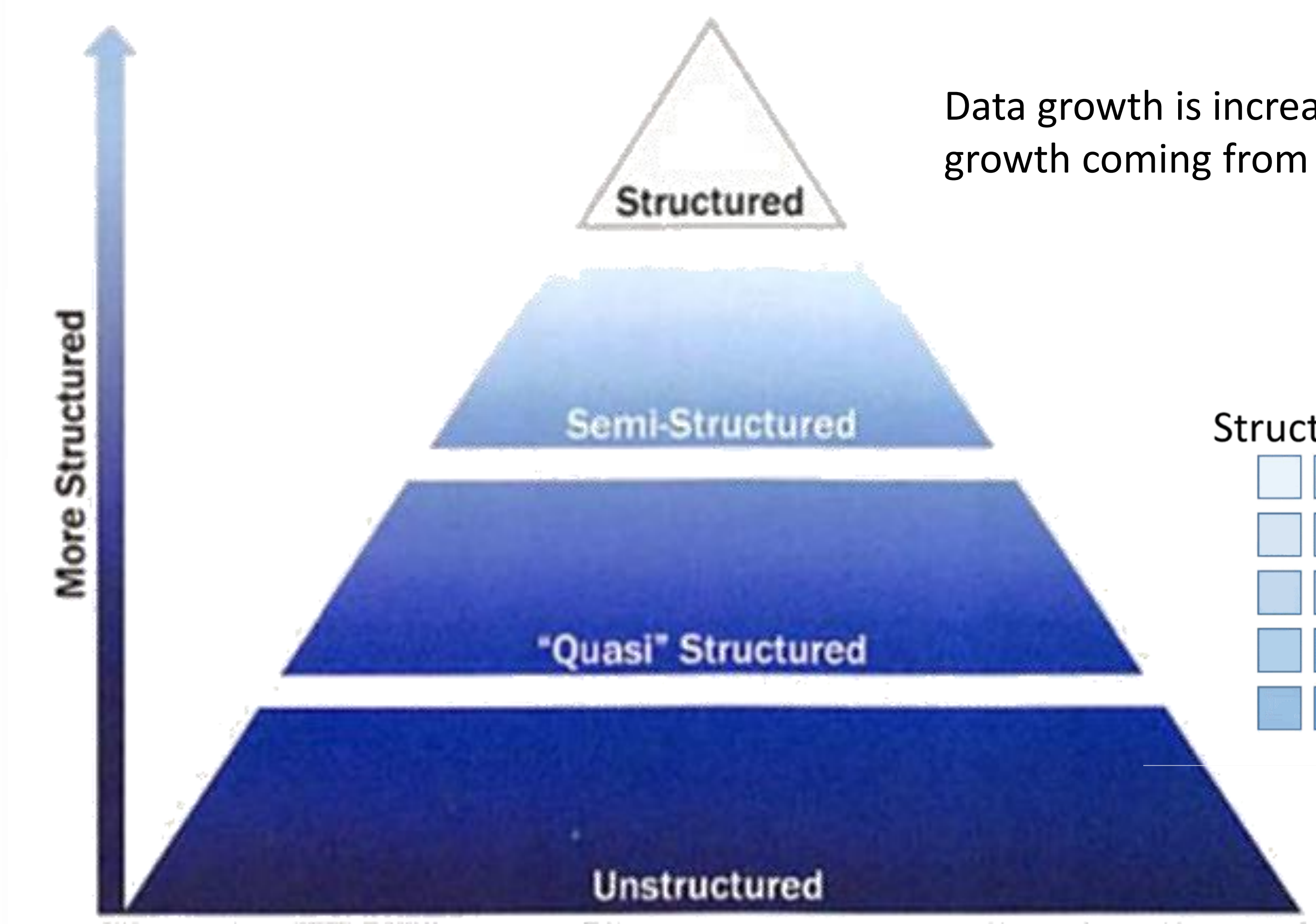
Data and the Data Ecosystem

SOURCES OF THE BIG DATA DELUGE

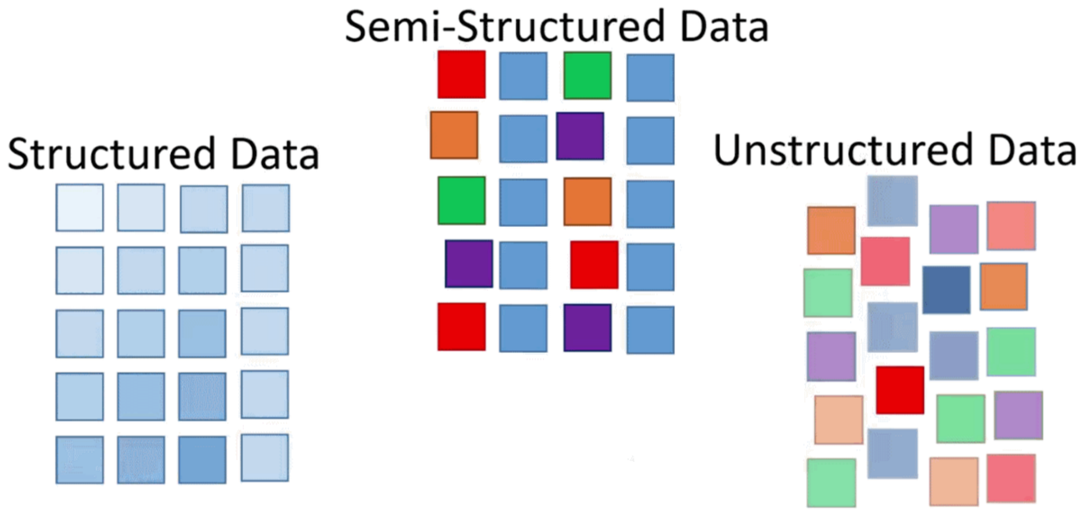


Data and the Data Ecosystem

DATA STRUCTURES



Data growth is increasingly unstructured, with 80 – 90% of future data growth coming from unstructured data types.



Data and the Data Ecosystem

DATA STRUCTURES

- Structured Data: contain a defined data type, format and structure

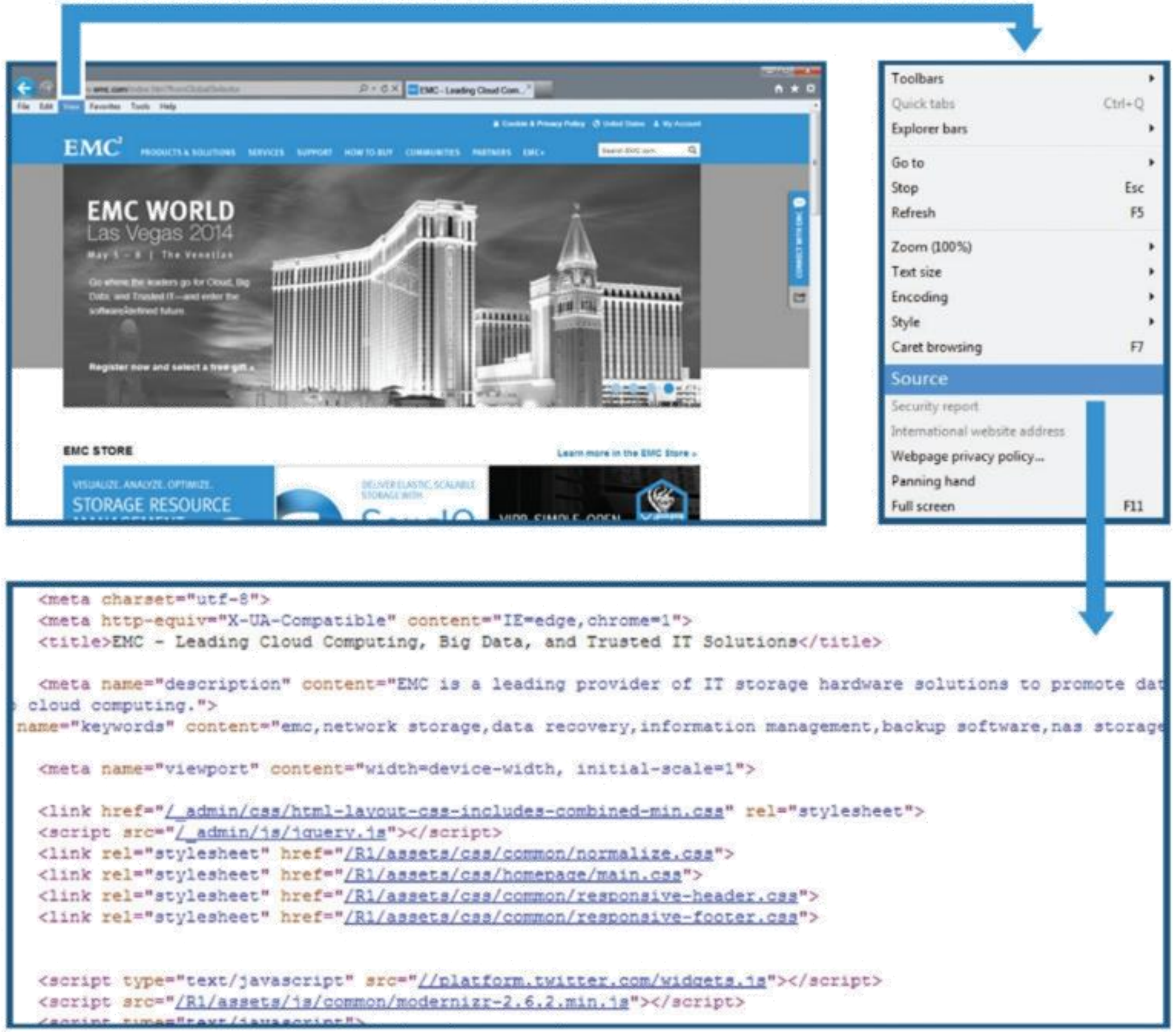
SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	------------Thousands------------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

Transaction data
Online Analytical Processing (OLAP) cubes
Traditional Relational Database Management Systems (RDBMS)
Comma-Separated Value (CSV) files
Simple spreadsheets

Data and the Data Ecosystem

DATA STRUCTURES

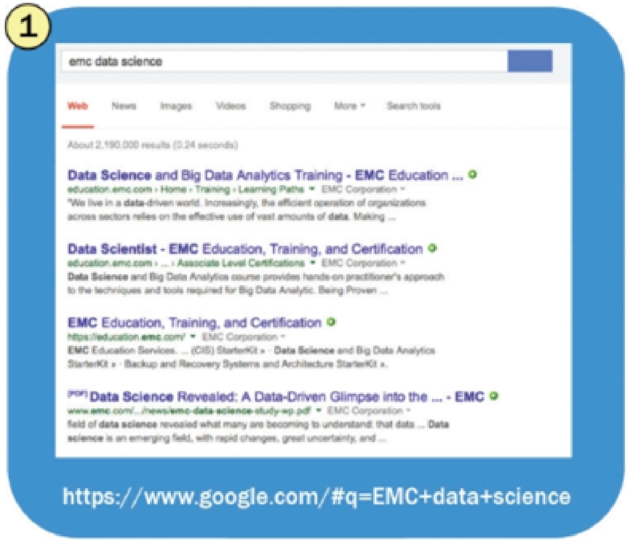
- Semi-Structured Data: textual data files with discernible pattern that enables parsing data files that are self-describing and defined by an Extensible Markup Language (XML) schema



Data and the Data Ecosystem

DATA STRUCTURES

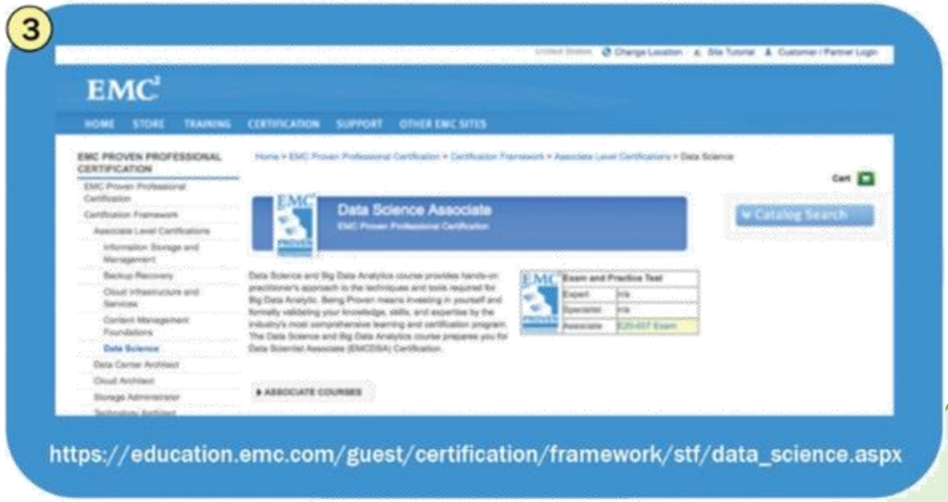
- Quasi-Structured Data: textual data with erratic data formats that can be formatted with effort, tools and time



https://www.google.com/#q=EMC+data+science

https://education.emc.com/guest/campaign/data_science.aspx

https://education.emc.com/guest/certification/framework/stf/data_science.aspx



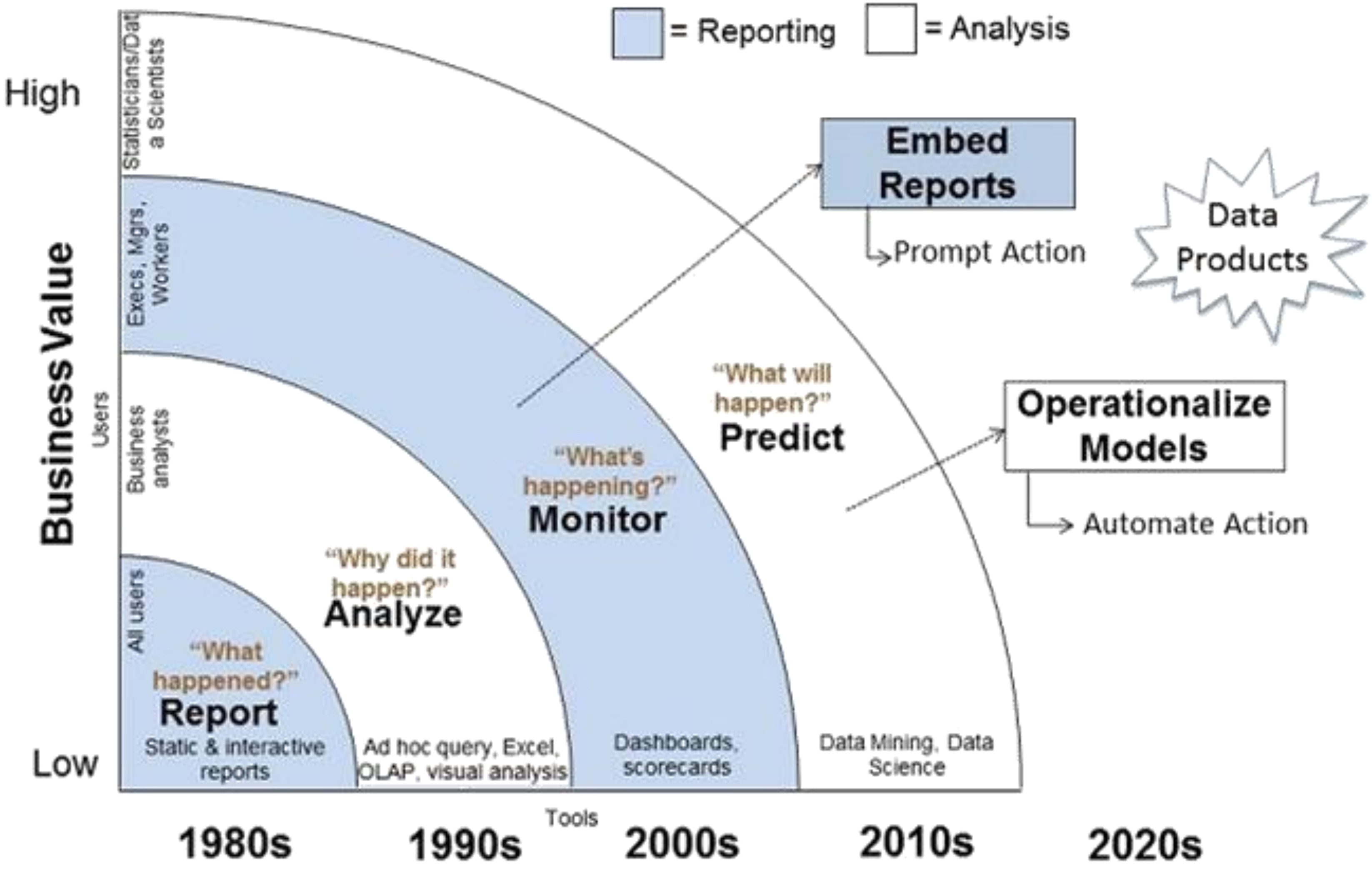
Data and the Data Ecosystem

BUSINESS DRIVERS FOR ADVANCED ANALYTICS

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-money laundering, fair lending

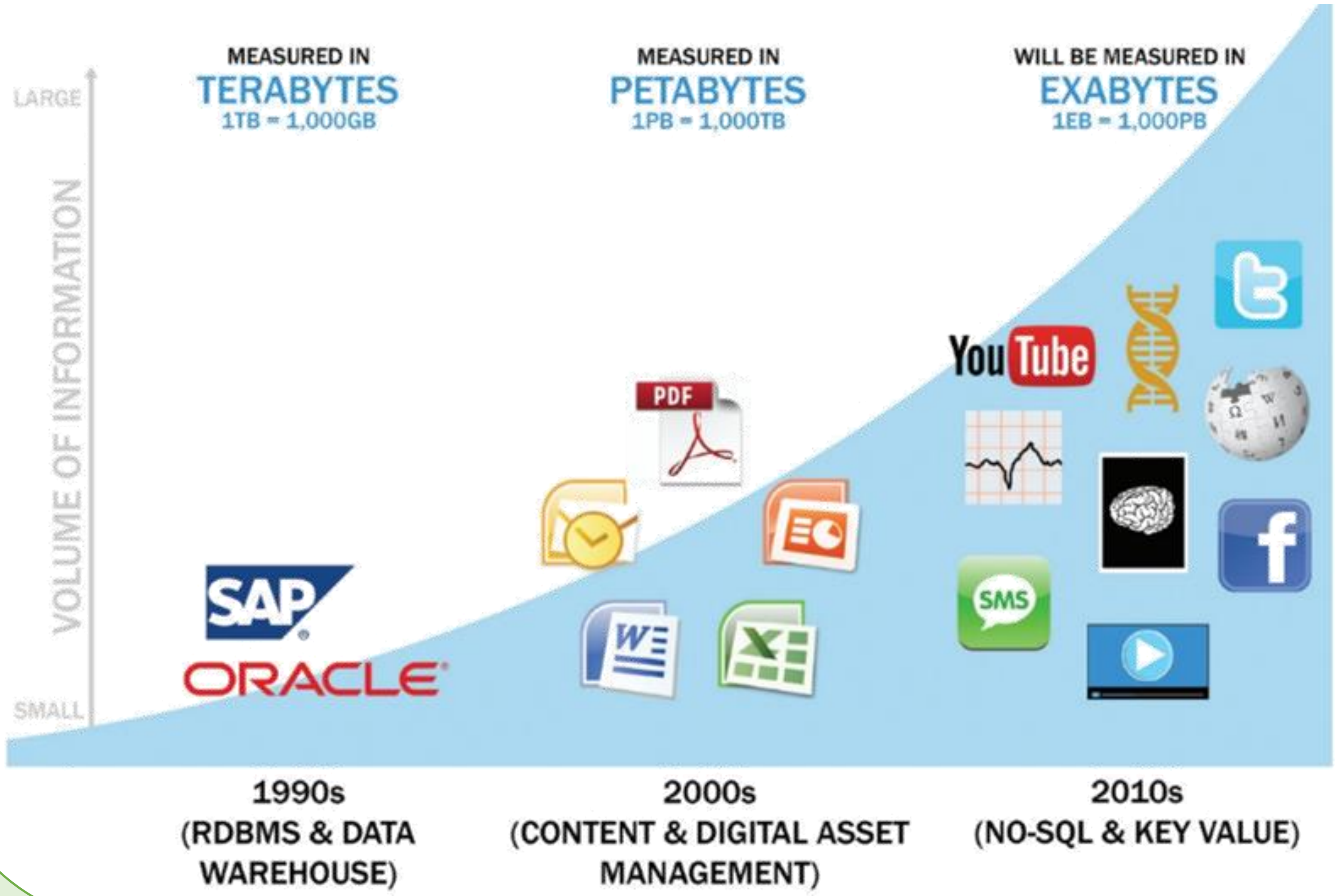
Data and the Data Ecosystem

EVOLUTION OF ANALYTICS



Data and the Data Ecosystem

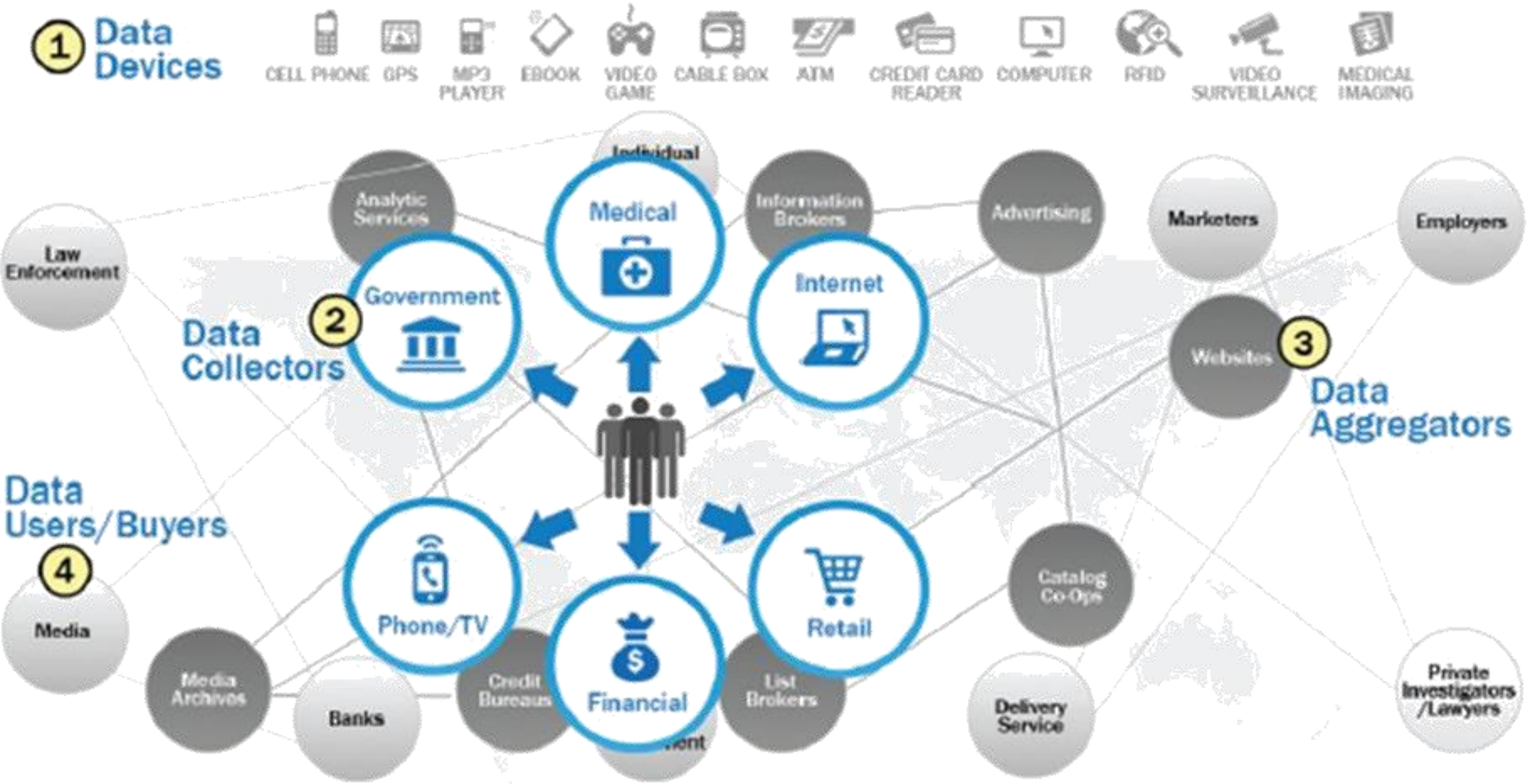
DRIVERS OF BIG DATA



- Medical information, such as genomic sequencing and diagnostic imaging
- Photos and video footage uploaded to the WW web
- Mobile devices – geospatial location data, metadata about text messages, phone calls, application usage on smart phones
- Smart devices
- Nontraditional IT devices – RFID readers, GPS navigation systems, seismic processing

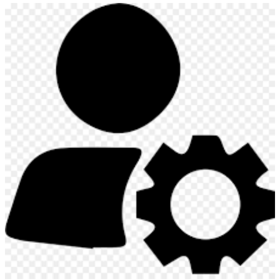
Data and the Data Ecosystem

THE BIG DATA ECOSYSTEM



Data and the Data Ecosystem

KEY ROLES IN THE BIG DATA ECOSYSTEM

Deep Analytical Talents	Data Savvy Professionals	Technology and Data Enablers
		
possess the skills to handle raw, unstructured data and to apply complex analytical techniques at massive scales	possess the basic knowledge of statistics or machine learning and can define key questions that can be answered using advanced analytics	provide technical expertise to support analytical projects, such as provisioning and administrating analytical sandboxes, and managing large-scale data architectures that enable widespread analytics within companies and other organizations
statisticians, economists, mathematicians	financial analysts, life scientists, operation managers, business managers	computer engineers, programmers, database administrators

Data Science and its Applications

Data Science is set of methodologies for taking in thousands of forms of available data and using them to draw meaningful conclusions.

It represents the optimization of processes and resources to produce data insights – data-informed conclusions or predictions that can be used to understand (and improve) health, businesses and investments, lifestyles and social lives.

MACHINE LEARNING

Case study: fraud detection

Amount	Date	Type	...
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

<u>Machine Learning Requisites</u>

- A well-defined question
  - *What is the probability that this transaction is fraudulent?*
- A set of example data
  - Old transaction labeled as “fraudulent” or “valid”
- A new set of data to use the algorithm on
  - New transactions

INTERNET OF THINGS (IoT)

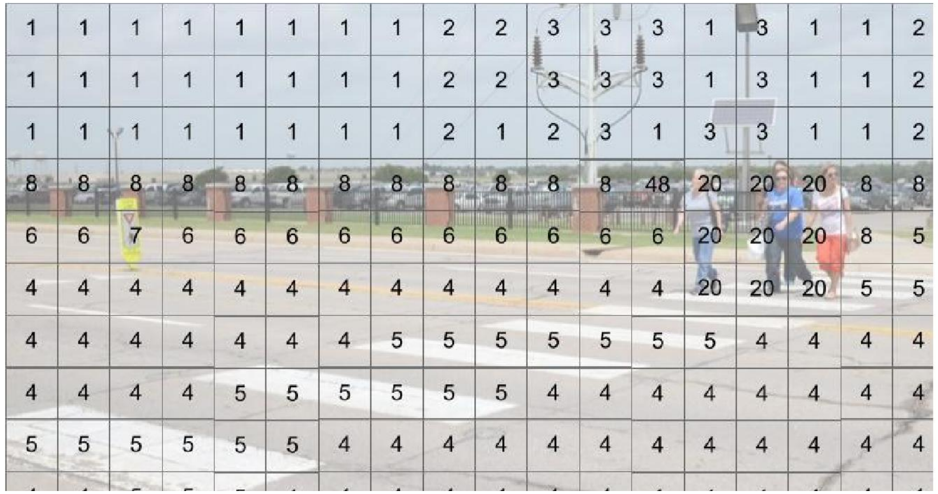
Case study: smart watch

- Gadgets that aren’t standard computers
  - Smart watches
  - Internet-connected home security systems
  - Electronic toll collection systems
  - Building energy management systems

Data Science and its Applications

DEEP LEARNING

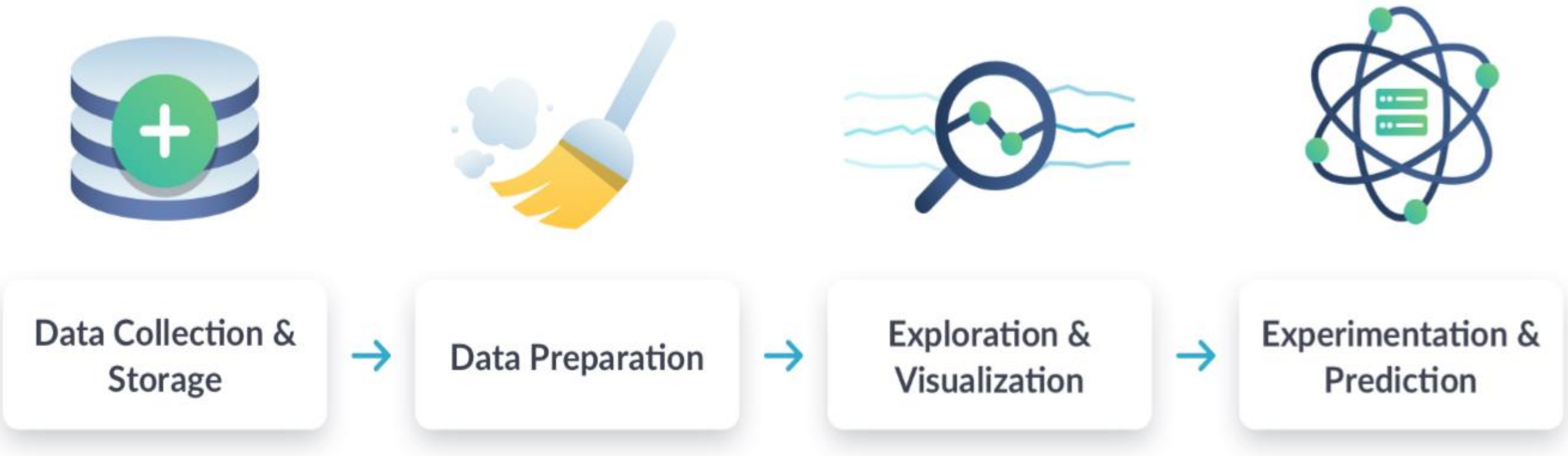
Case study: image recognition



- Many neurons working together
- Requires more training data
- Used in complex problems
    - Image classification
    - Language learning / understanding

Data Science Roles and Tools

THE DATA SCIENCE WORKFLOW

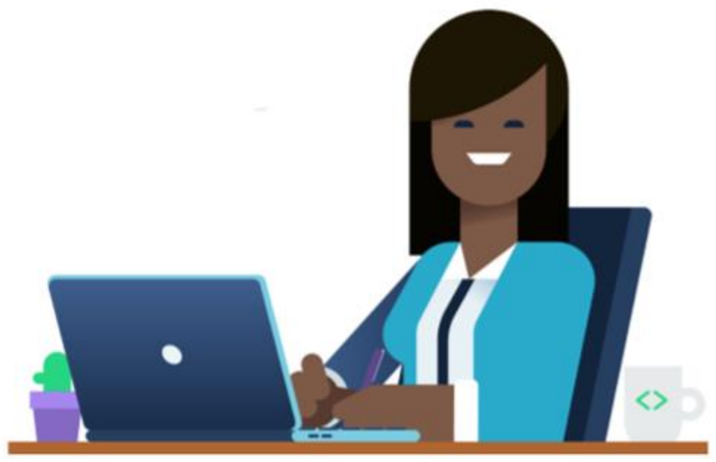


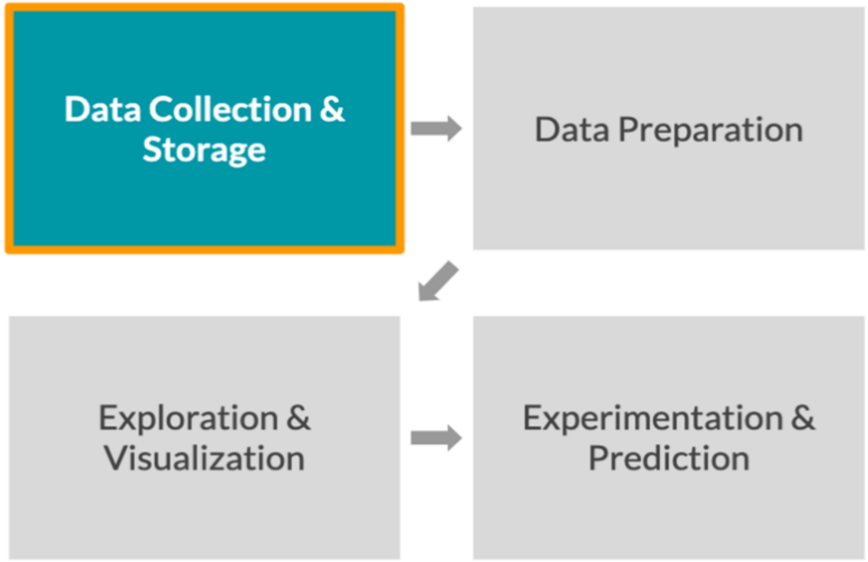
Surveys Web results Geo-tagged posts Financial transactions	Finding missing or duplicate values Converting data	Building dashboards Comparing data	Building sytems Validating sytems Performing tests

Data Science Roles and Tools

DATA ENGINEER

- Information architects
- Control the flow of data
- Build data pipelines and storage solutions so that data is easily collected, obtained and processed



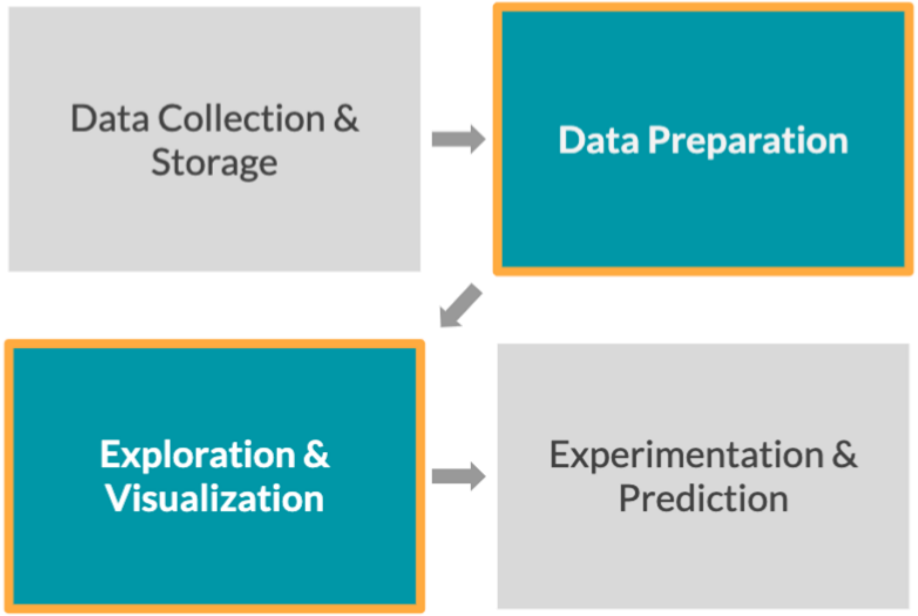


- SQL: to store and organize data
- Java, Scala or Python: programming languages to process data
- Shell: command line to automate and run tasks
- AWS, Azure, Google Cloud Platform: cloud computing to ingest and store large amounts of data

Data Science Roles and Tools

DATA ANALYST

- Perform simple analyses to describe data
- Explore and clean the data and create visualizations and dashboards to summarize data
- Describe the present via data

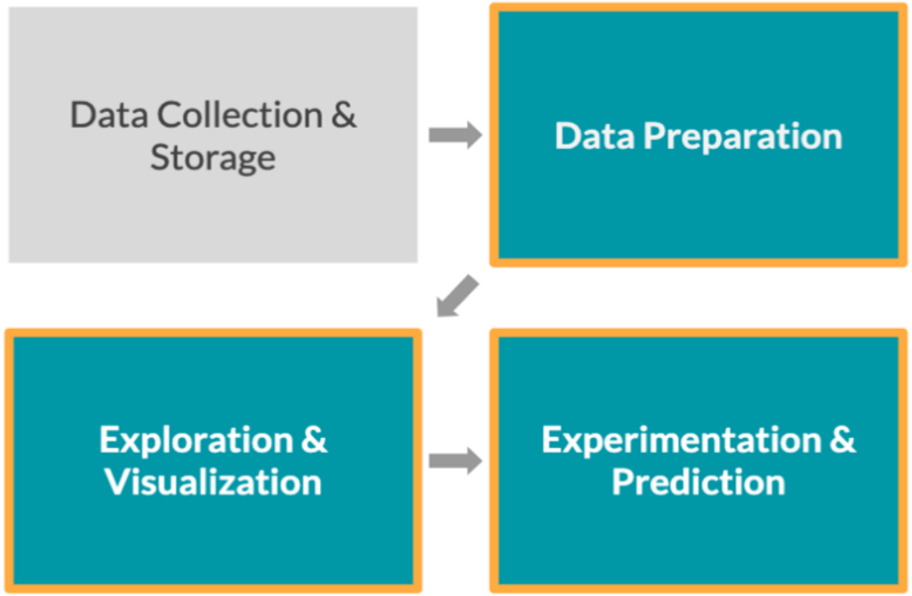


- SQL: to retrieve and aggregate data relevant to the analysis
- Spreadsheets (Excel or Google Sheets): to perform simple analyses on small quantites of data
- BI Tools (Tableau, Power BI, Looker): to create dashboards and share analyses
- Python, R: for cleaning and analyzing data

Data Science Roles and Tools

DATA SCIENTIST

- Find new insights from data from statistical data, rather than solely describing data
- Use traditional machine learning for prediction and forecasting

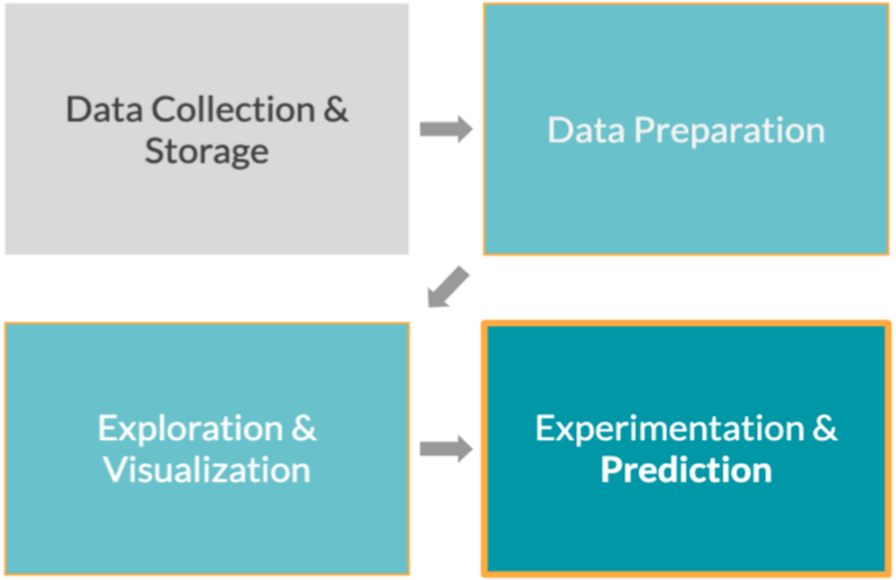


- SQL: to retrieve and aggregate data relevant to the analysis
- Python and/or R with associated libraries (pandas / tidyverse)

Data Science Roles and Tools

MACHINE LEARNING SCIENTIST

- Similar to data scientists, but with a machine learning specialization
- Go beyond machine learning with deep learning
- Strong focus on prediction



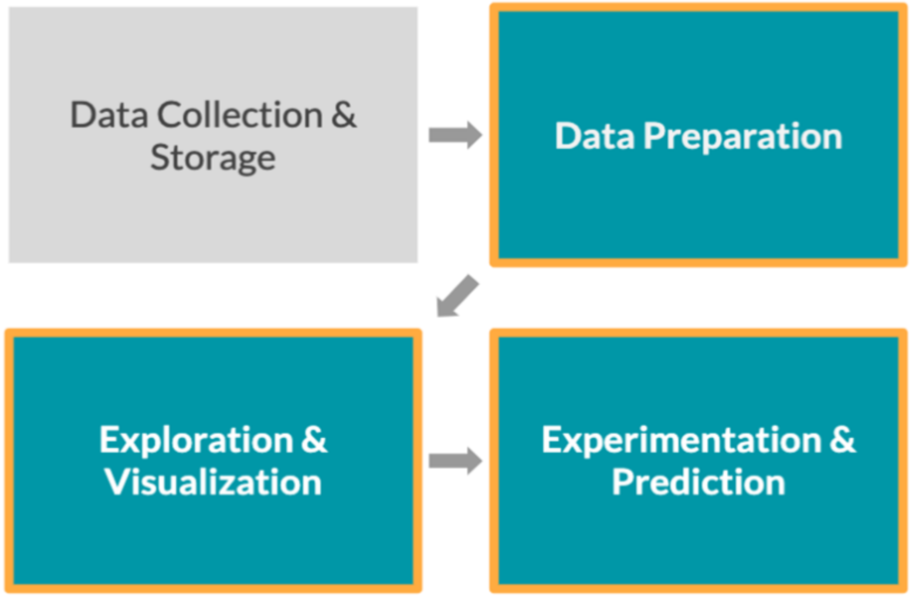
- R, Python: to create predictive models, with associated libraries, like TensorFlow to run powerful deep learning algorithms

Data Science Roles and Tools

Data Engineer	Data Analyst	Data Scientist	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Gain insights from data	Predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R	Python/R

The Data Scientist

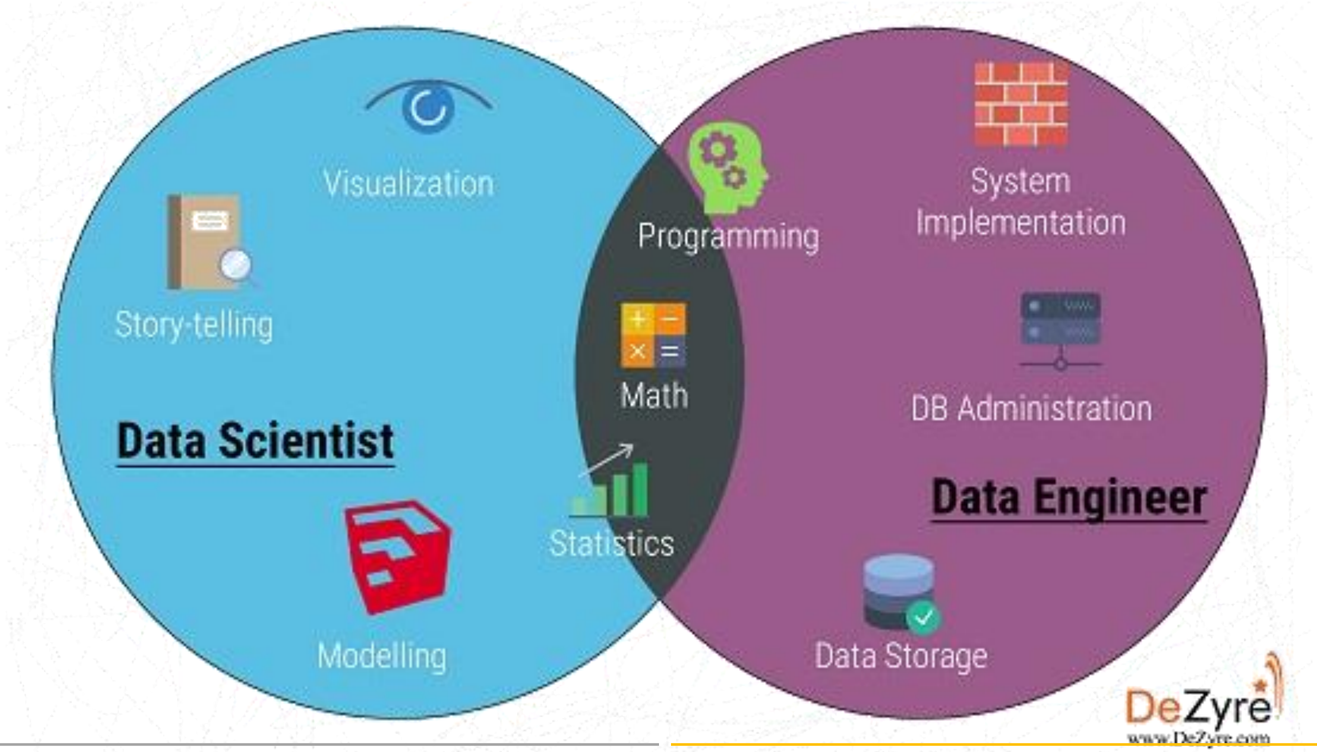
- Find new insights from data from statistical data, rather than solely describing data
- Use traditional machine learning for prediction and forecasting



- SQL: to retrieve and aggregate data relevant to the analysis
- Python and/or R with associated libraries (pandas / tidyverse)

The Data Scientist

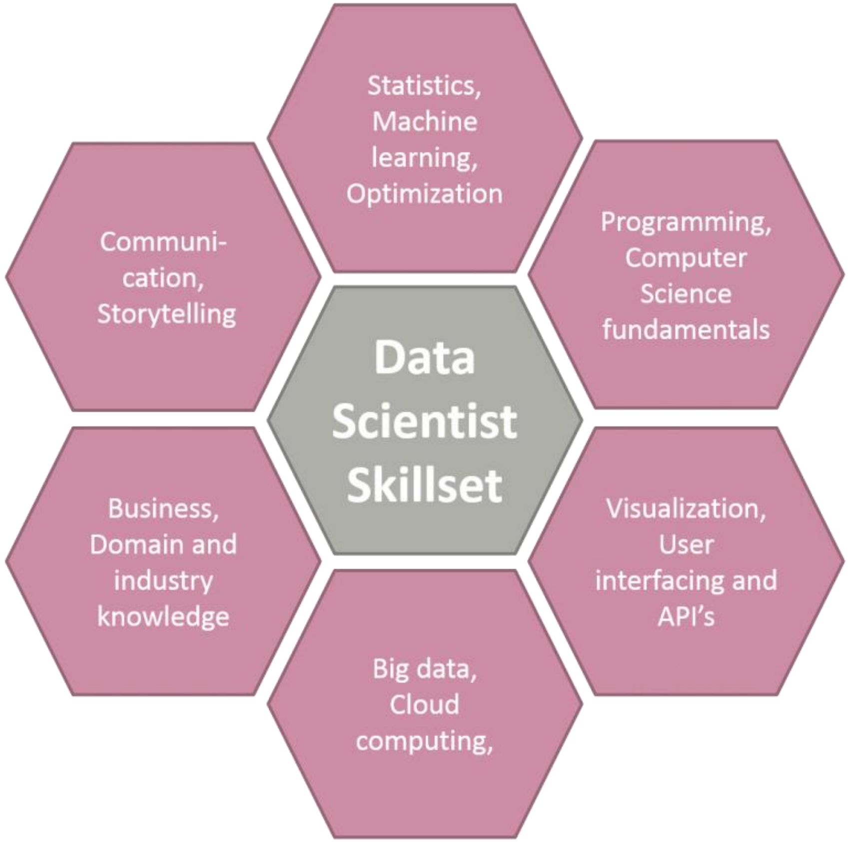
THE DATA SCIENTIST VS THE DATA ENGINEER

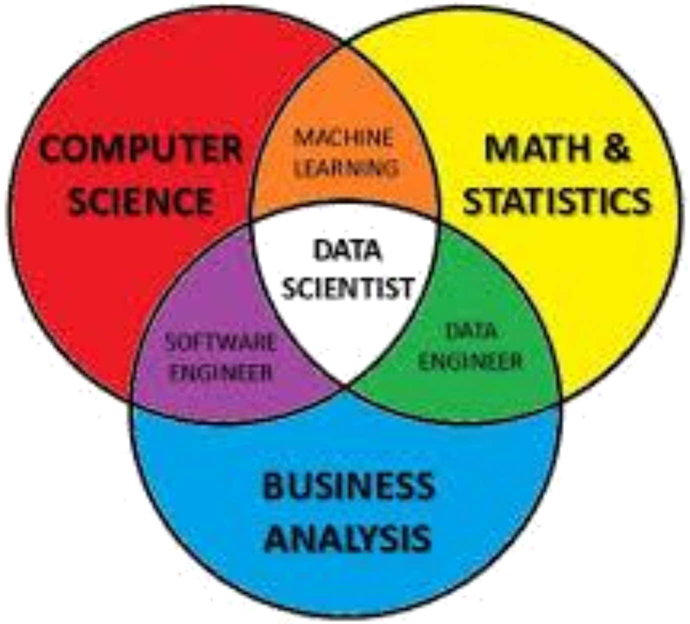


DATA SCIENCE	DATA ENGINEERING
the computational science of extracting meaningful insights from raw data and then effectively communicating those insights to generate value	the engineering domain that is dedicated to building and maintaining systems that overcome data processing bottlenecks and data handling problems

The Data Scientist

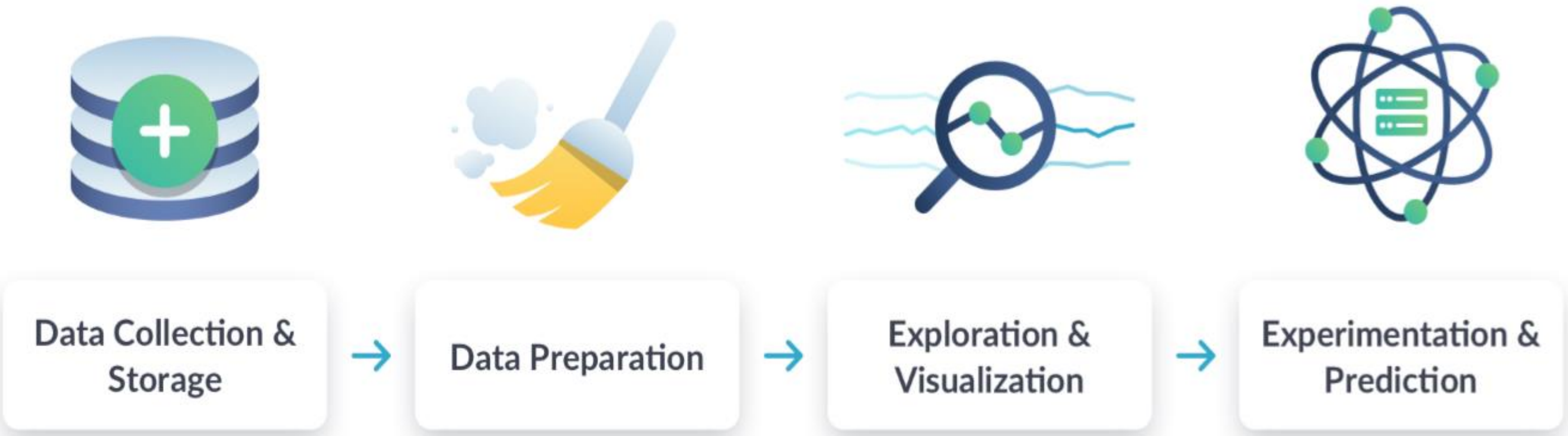
DATA SCIENTISTS' SKILL SET AND BEHAVIORAL CHARACTERISTICS





The Data Science Workflow

THE DATA SCIENCE WORKFLOW



Surveys
Web results
Geo-tagged posts
Financial transactions

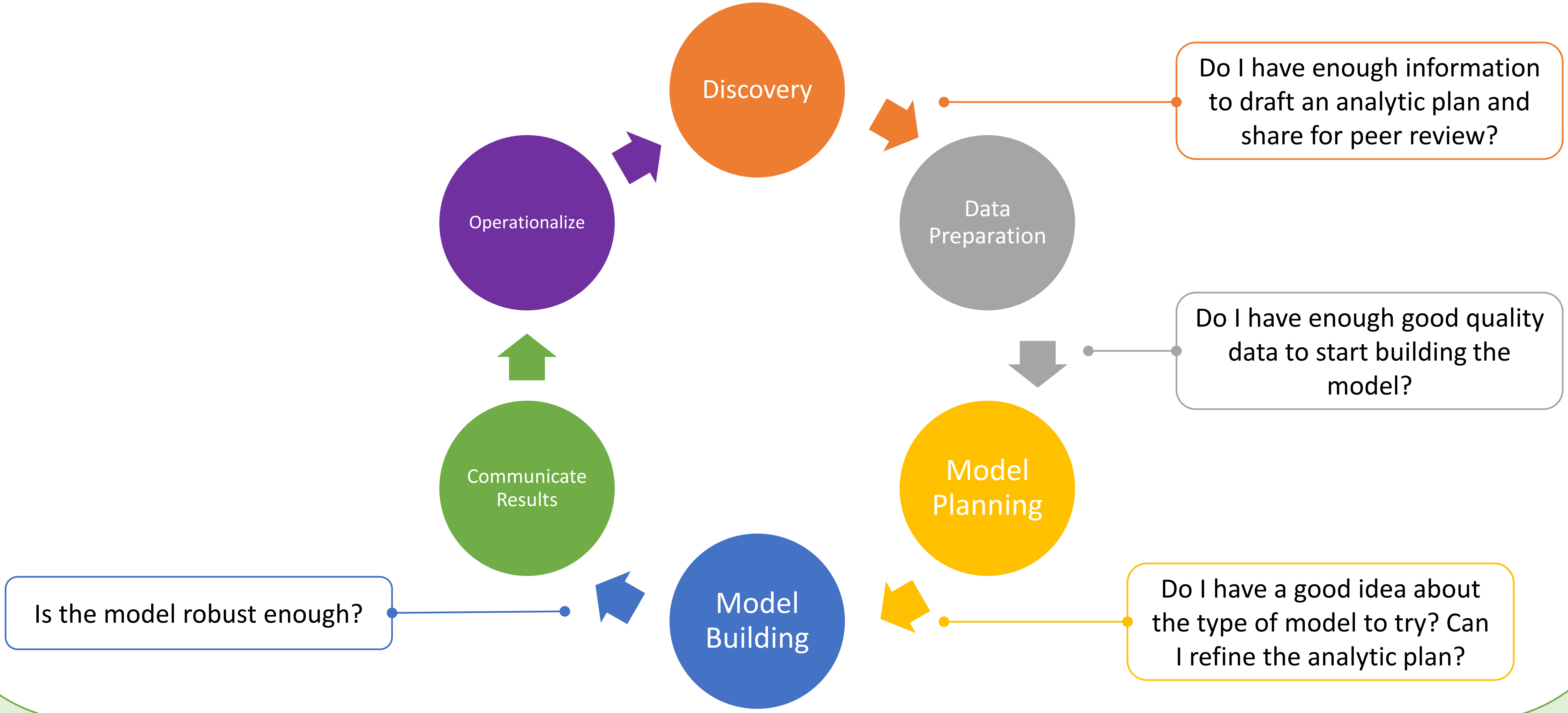
Finding missing or
duplicate values
Converting data

Building dashboards
Comparing data

Building sytems
Validating sytems
Performing tests

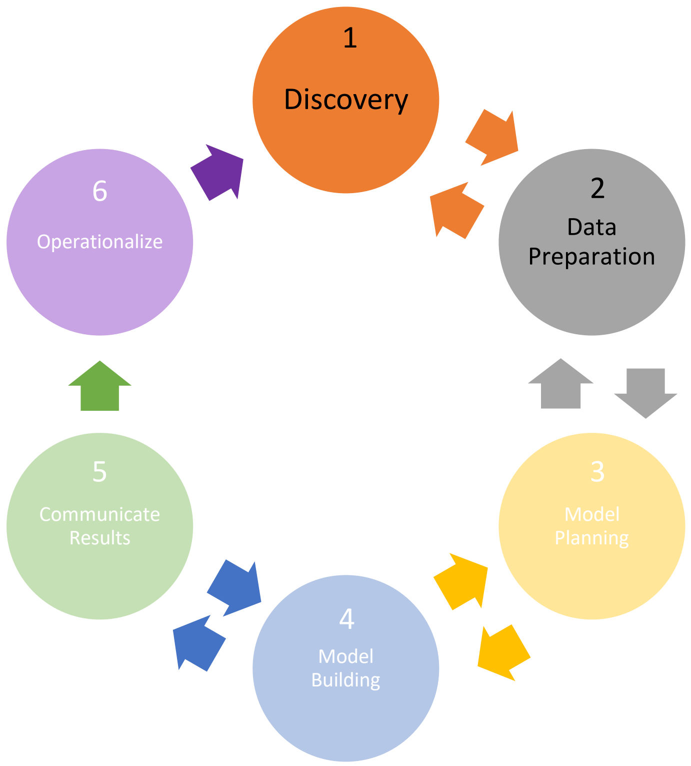
The Data Science Workflow

THE ANALYTICS LIFECYCLE



The Data Science Workflow

THE ANALYTICS LIFECYCLE



The team learns the business domain, including relevant history such as whether the unit has attempted similar projects in the past from which they can learn.

The team assesses the resources available to support the project in terms of people, technology, time and data.

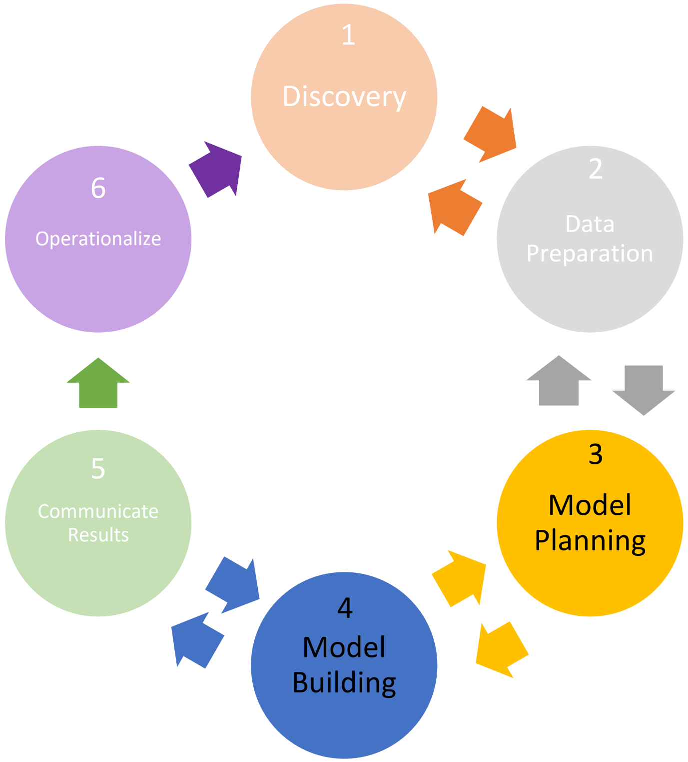
The team frames the business problem as an analytics challenge that can be addressed in subsequent phases and formulates initial hypotheses (IHs) to test and begin learning the data.

The team executes extract, load and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox so the team can work with it and analyze it.

The team familiarizes itself with the data thoroughly and take steps to condition it.

The Data Science Workflow

THE ANALYTICS LIFECYCLE



The team determines the methods, techniques and workflow it intends to follow for the subsequent model building phase.

The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

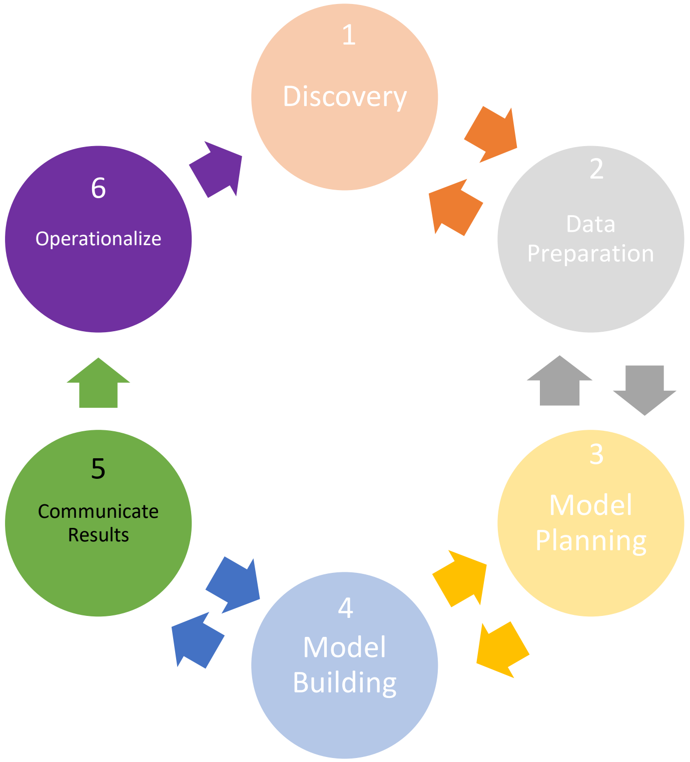
The team develops datasets for testing, training and production purposes.

The team builds and executes models based on the work done in the model planning phase.

The team considers whether its existing tools will suffice for running the models, or if it will need a more robust environment fro executing models and workflows

The Data Science Workflow

THE ANALYTICS LIFECYCLE



The team determines if the results of the project are a success or a failure based on the criteria developed in Phase 1.

The team identifies key findings, quantifies the business value, and develops a narrative to summarize and conveys findings to stakeholders.

The team delivers final reports, briefings, code, and technical documents.

The team may run a pilot project to implement the models in a production environment.

The Data Science Workflow

THE ANALYTICS LIFECYCLE: DISCOVERY

Start

↓

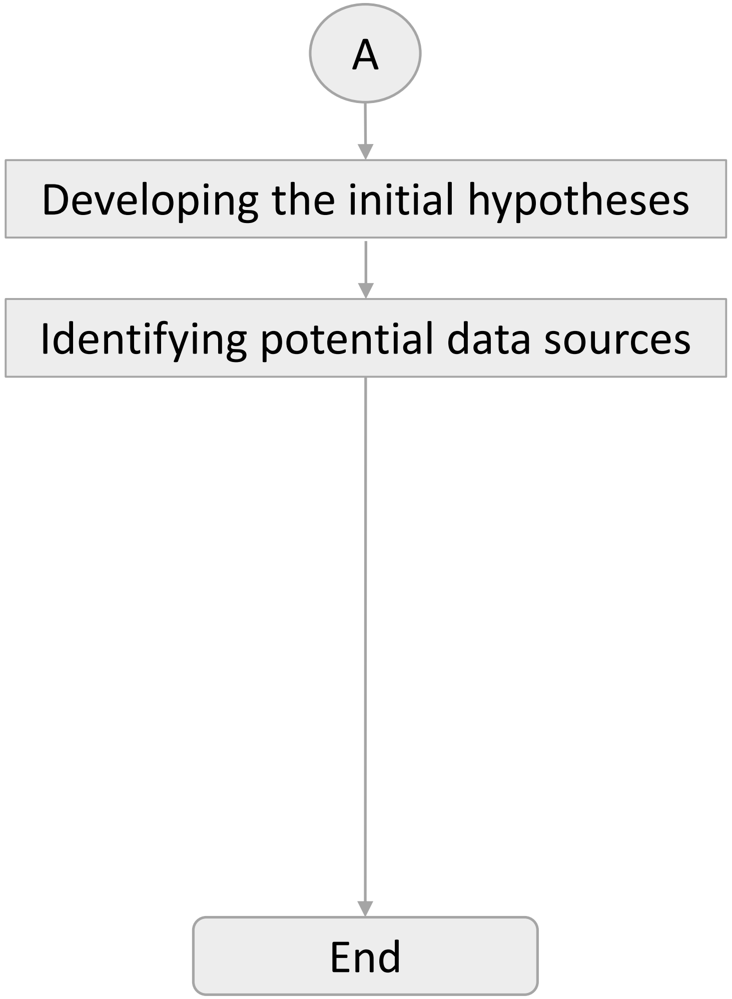
Learning the business domain	How much business domain or knowledge is needed to develop models?
Identifying resources	What technology, tools, systems, data, people are needed?
Framing the Problem	What is the problem, the objectives and success and failure criteria?
Identifying Key Stakeholders	Who will benefit from the project or will be significantly impacted by the project?
Interviewing the Analytics Sponsor	Who has final decision-making authority on the project? How will the focus and scope of the problem change if the following dimensions change: time, people, risk, resources, size and attributes of data?

↓

A

The Data Science Workflow

THE ANALYTICS LIFECYCLE: DISCOVERY



What are the ideas that the team can test with data?

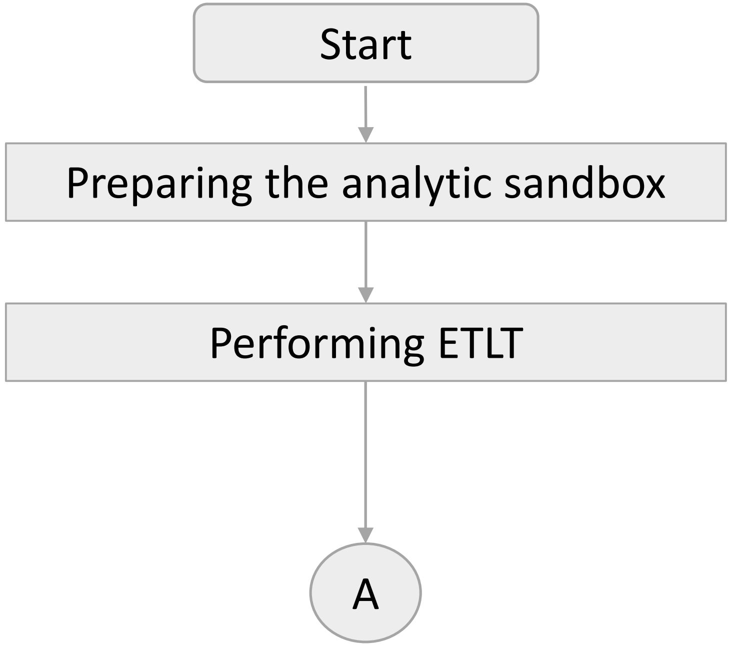
What data (including volume, type and time span) will the team need to solve the problem?

Activities:

- Identify data sources: inventory of available and needed datasets
- Capture aggregate data sources
- Review the raw data: quality and limitations of data
- Evaluate the data structures and tools needed
- Scope the sort of data infrastructure needed for this type of problem: disk storage and network capacity

The Data Science Workflow

THE ANALYTICS LIFECYCLE: DATA PREPARATION



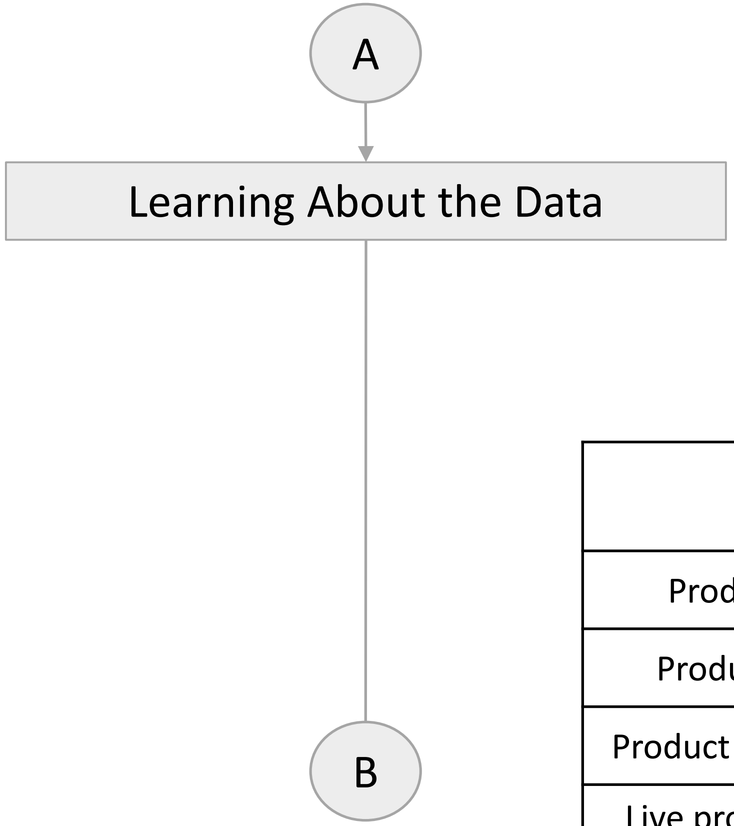
obtaining an analytic sandbox (or workspace) in which the team can explore the data without interfering with live production databases

assessing data quality and structuring the datasets properly so they can be used for robust analysis

extracting, transforming, loading or extracting, loading, transforming data (advocated by analytic sandboxes to preserve raw data ~ a good practice is to make an inventory of the raw and current data available to datasets)

The Data Science Workflow

THE ANALYTICS LIFECYCLE: DATA PREPARATION

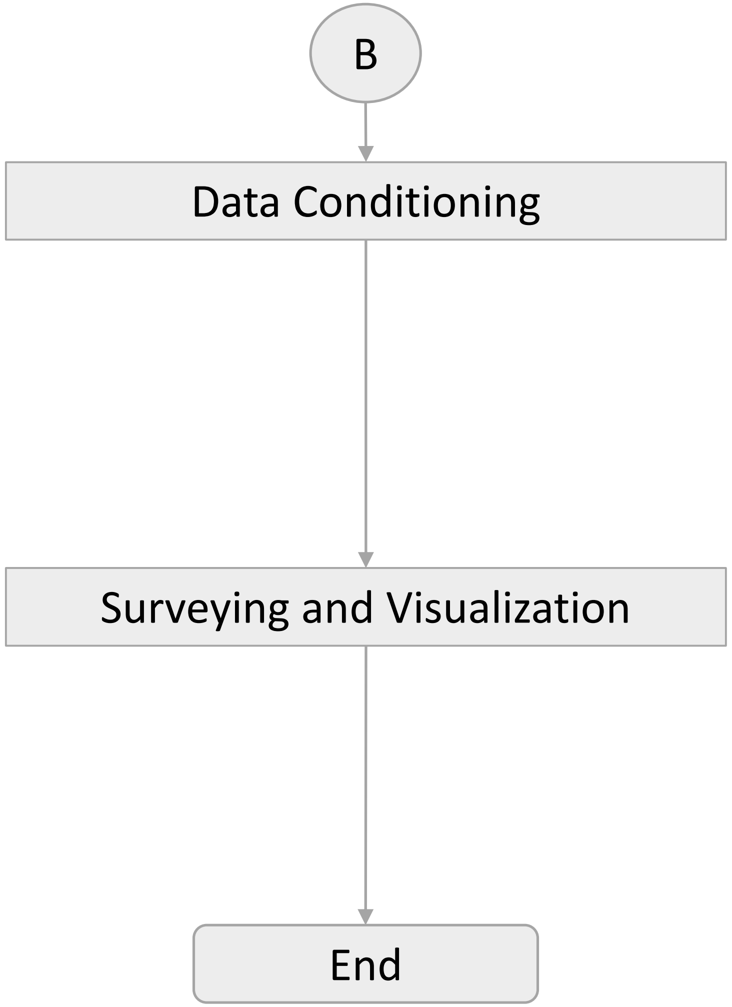


- understanding what constitutes a reasonable value and expected output versus what is a surprising finding
- cataloging the data sources that the team has access to and identifying additional sources that the team can leverage but does not have access to

Dataset	Data Available and Accessible	Data Available but Not Accessible	Data to Collect	Data to Obtain from 3rd Party Sources
Product shipped	✓			
Product financials		✓		
Product call center data		✓		
Live product feedback surveys			✓	
Product sentiment from social media				✓

The Data Science Workflow

THE ANALYTICS LIFECYCLE: DATA PREPARATION



determining the data sources and how clean the data is
determining to what degree the data contains missing or inconsistent values and if the data contains values that deviate from normal
assessing the consistency of the data types
reviewing data content
looking for evidence of systematic error

leveraging data visualization tools to gain an overview of the data ~ examining data quality for unexpected values or skewness

Guidelines:

- Ensure that calculations are consistent and that data distribution is consistent
- Assess the granularity and aggregation of the data and the range of values
- Determine whether the data is standardized / normalized

The Data Science Workflow

THE ANALYTICS LIFECYCLE: MODEL PLANNING

Start

Data Exploration and Variable Selection

- understanding the relationships among the variables to inform selection of the variables and methods and to understand the problem domain (a good way is to use tools for data visualization)
- testing a range of variable to include in the model and then focusing on the most important and influential variables

Model Selection

- choosing an analytical technique, or a short list of candidate techniques, based on the end goal of the project

End

The Data Science Workflow

THE ANALYTICS LIFECYCLE: MODEL BUILDING

<u>Considerations</u>

- Does the model appear valid and accurate on the test data?
- Does the model output / behavior make sense to the domain experts?
- Do the parameter values of the fitted model make sense in the context of the domain?
- Is the model sufficiently accurate to meet the goal?
- Does the model avoid intolerable mistakes? ~ false positives and false negatives
- Are more data or more inputs needs? Do any of the inputs need to be transformed or elimitated?
- Will the kind of model chosen support the runtime requirements?
- Is a different form of the model required to address the business problem?

The Data Science Workflow

THE ANALYTICS LIFECYCLE: RESULTS COMMUNICATION

- Start
- Record all findings; select the three most significant to share with the stakeholders
- Reflect on the implications of the findings
- Measure the business impact of the results and demonstrate the value of the findings
- Make recommendations for future work or improvements to existing processes
- End

The Data Science Workflow

THE ANALYTICS LIFECYCLE: OPERATIONALIZATION

Start

↓

Deploy the new analytical methods or models in a production environment (pilot level)

↓

Create a mechanism for perform ongoing monitoring of model accuracy

↓

Learn from the deployment and make the needed adjustments

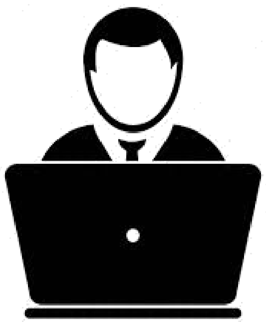
↓

End

The Data Science Workflow

KEY ROLES FOR A SUCCESSFUL ANALYTICS PROJECT

Business User



Understands the domain area and usually benefits from the results

Consults and advises the project team on the context of the project, the value of the results and how the outputs will be operationalized

Project Sponsor

Provides the requirements for the project and defines the core business problem

Provides the funding and gauges the degree of value from the final outputs of the working team, sets priorities for the project and clarifies the desired outputs

Project Manager

Ensures that key milestones and objectives are met on time and at the expected quality

Business Intelligence Analyst

Provides business domain expertise based on an understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective

Create dashboards and reports and have knowledge of the data feeds and sources

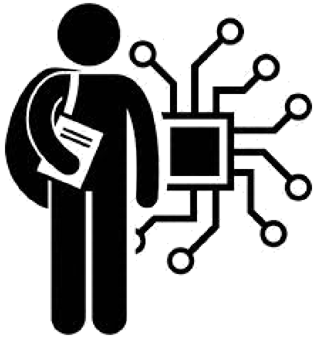
The Data Science Workflow

KEY ROLES FOR A SUCCESSFUL ANALYTICS PROJECT

Database Administrator

- Provisions and configures the database environment to support the analytics needs of the working team
- Provides access to key databases or tables and ensures the appropriate security levels are in place related to the data repositories

Data Engineer



- Assists with turning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox
- Executes the actual data extractions and performs substantial data manipulation to facilitate the analytics

Data Scientist

- Provides subject matter expertise for analytical techniques, data modeling and applying valid analytical techniques to given business problems
- Ensures overall analytics objectives are met
- Designs and executes analytical methods and approaches with the data available to the project

The Data Science Workflow

ANALYTICS DELIVERABLES

Presentation for Project Sponsors

Contains high-level takeaways for executive level stakeholders, with a few key messages to aid their decision-making process; contains clean, easy visuals for the viewer to grasp

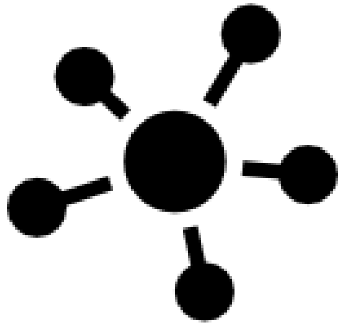
Presentation for Analysts

Describes business process changes and reporting changes; contains details and technical graphs

Code

For technical people

Technical Specifications



For implementing the code

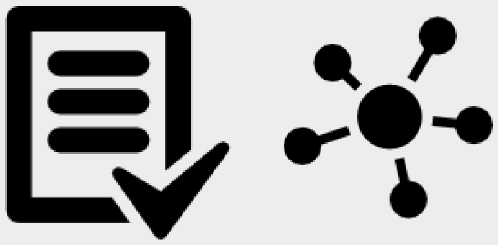
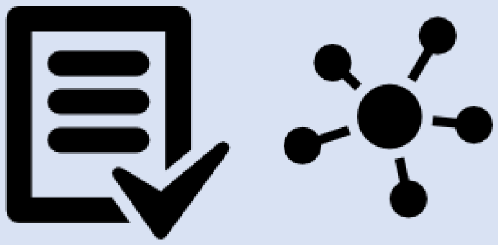
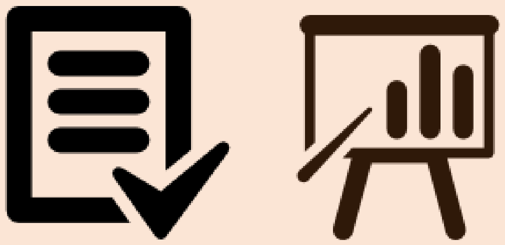
The Data Science Workflow

PROJECT STAKEHOLDERS' OUTPUTS

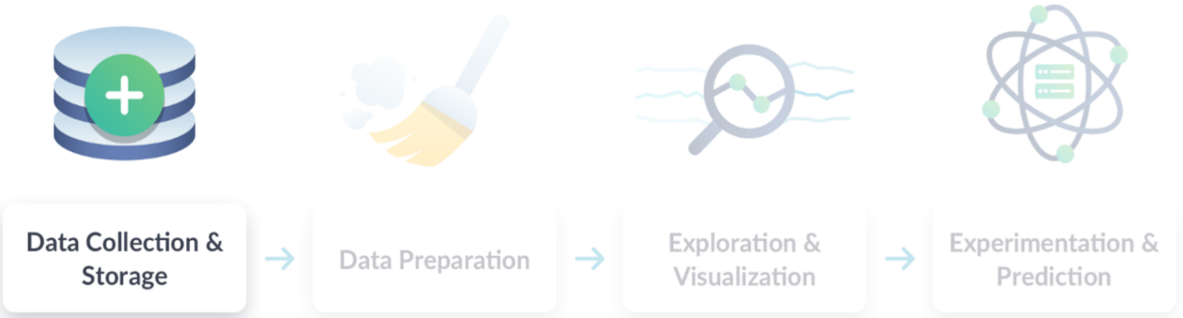
Business User	Project Sponsor	Project Manager	Business Intelligence Analyst
determines the benefits and implications of the findings to the business	asks questions related to the business impact of the project, the risks and return on investment, and the way the project can be evangelized within the organization	determines if the project was completed on time and within budget and how well the goals were met	determines if the reports and dashboards he manages will be impacted and need to change

The Data Science Workflow

PROJECT STAKEHOLDERS’ OUTPUTS

Database Administrator	Data Engineer	Data Scientist
		
		
needs to share his code from the analytics project and create a technical document on how to implement it	needs to share his code from the analytics project and create a technical document on how to implement it	needs to share the code and explain the model to his peers, managers and other stakeholders

Data Collection and Storage



DATA SOURCES

Company Data

- Web events
- Survey data
- Customer data
- Logistics data
- Financial transactions

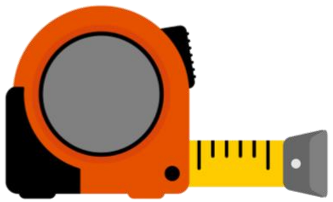
Open Data

- Application Programming Inferfaces
    - Twitter
    - Wikipedia
    - Google Maps
- Public Records
    - International Organizations (World Bank)
    - National Statistical Offices (surveys)
    - Government Agencies (weather, population)

Data Collection and Storage

DATA TYPES

Quantitative Data



- can be counted, measured and expressed using numbers

Qualitative Data

- can be observed but not counted
- descriptive and conceptual

60 inches tall
has 2 apples
costs $1,000

red
made in Italy
smells like fish

Data Collection and Storage

DATA TYPES

Image Data

- made up of pixels that contain information about color and intensity

Text Data

"Great evening, extremely good value"

Review of L'Ange 20 Restaurant

I went to this place with my boyfriend for a special occasion and we were not disappointed. We were greeted warmly by Christopher who guided us through the menu and wine. The food was delicious and I only wish that we could have had room for three courses. The value was excellent compared to other prices we had seen and we found the quality/value and atmosphere hard to match during the rest of our stay.

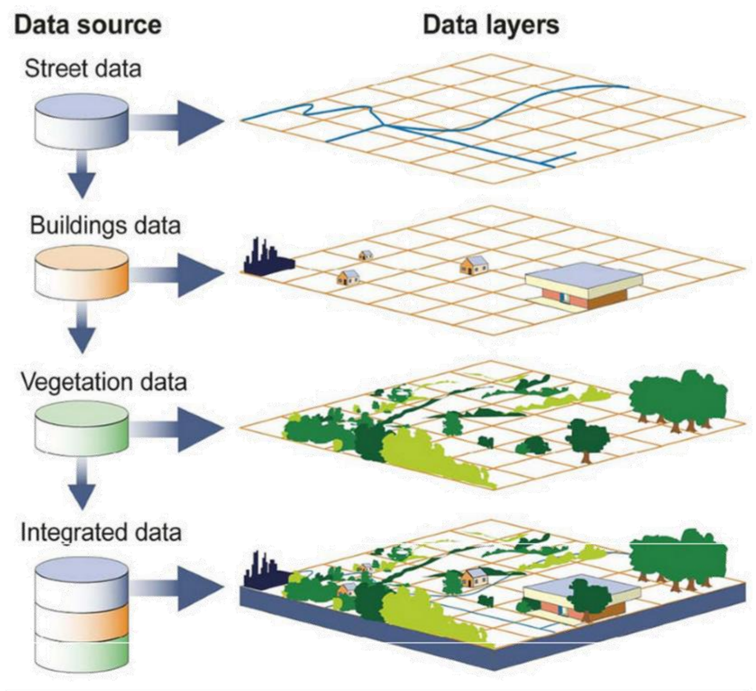
I had the lamb which I can highly recommend. When we return to Paris we will go back!

- social media posts
- reviews
- emails
- documents

Data Collection and Storage

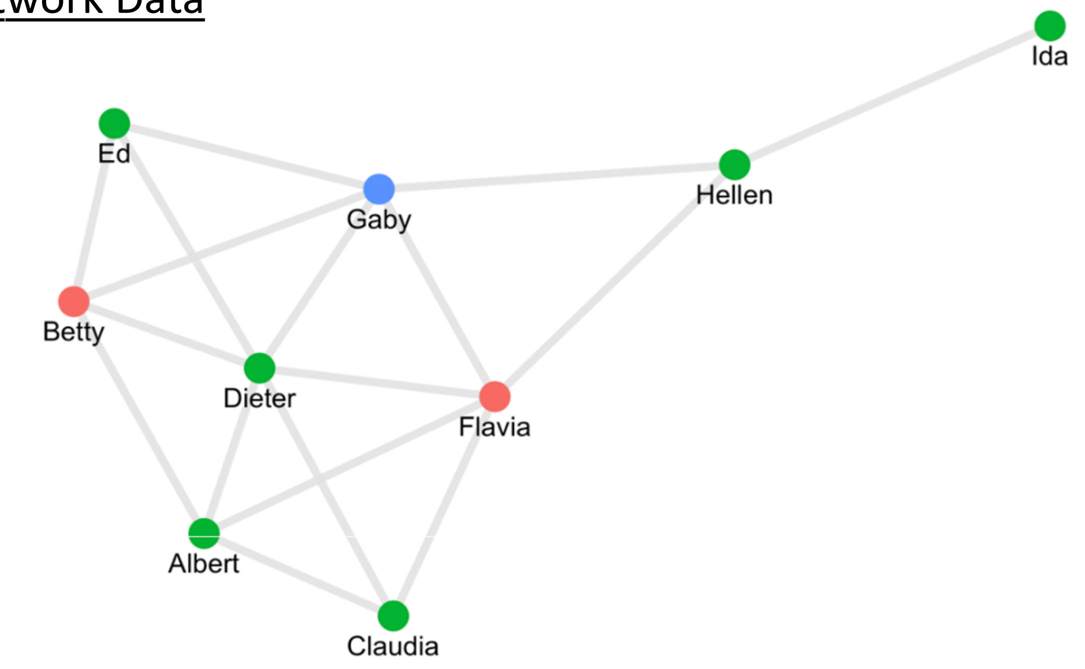
DATA TYPES

<u>Geospatial Data</u>



- data with location information (i.e., roads, buildings, vegetation)
- useful for navigation apps

<u>Network Data</u>



- shows relationships between nodes (people or things)

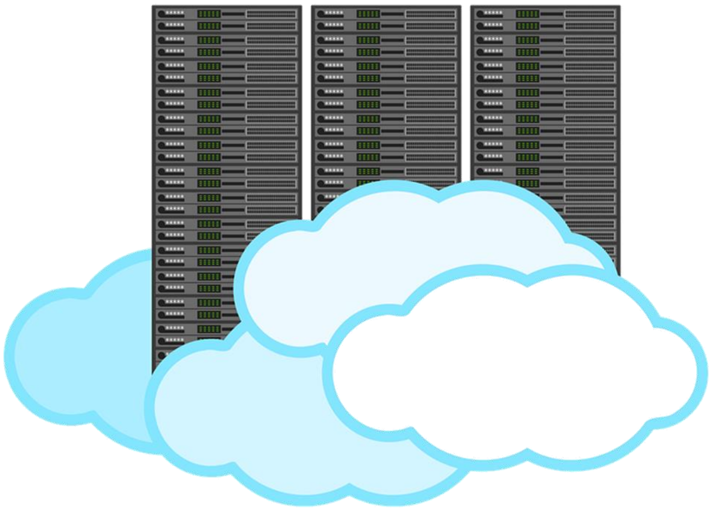
Data Collection and Storage

DATA STORAGE AND RETRIEVAL

Location: Parallel Storage

- cluster or servers for storage and easy access to data

Location: Cloud Storage



- Microsoft Azure
- Amazon Web Services (AWS)
- Google Cloud

Data Collection and Storage

DATA STORAGE AND RETRIEVAL

Type: Document Database

- stores unstructured data (i.e, email, text, video and audio files, social media messages)

Type: Relational Database

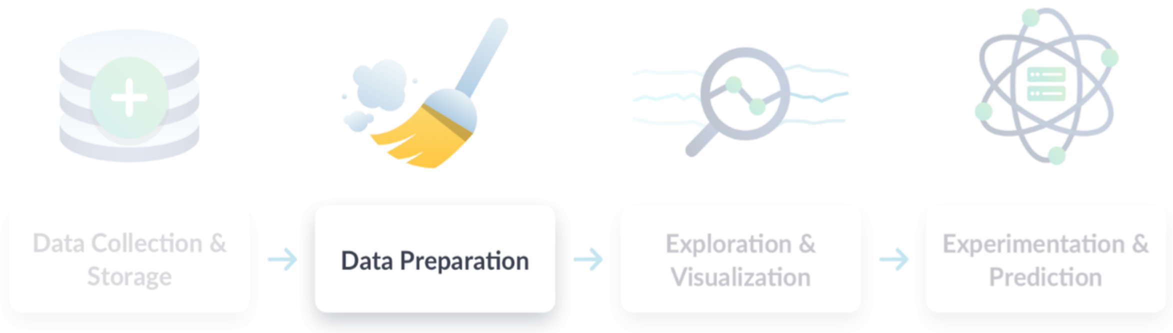
- stores structured data

Customer Name	Customer Address	...
Jane Doe	123 Maple St.	...

Retrieval: Data Query

Data Type	Query Language
Document Database	NoSQL
Relational Database	SQL

Preparation, Exploration and Visualization



DATA PREPARATION

<u>Why?</u>

- Real-life data is messy
- To prevent:
  - errors
  - incorrect results
  - biasing algorithms

	Sara	Lis	Hadrien	Lis
Age	“27”	“30”		“30”
Size	1.77	5.88	1.80	5.58
Country	“Belgium”	“USA”	“FR”	“USA”

Preparation, Exploration and Visualization

Tidy Data

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.88	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

- The observations (people) are in columns and their features are on rows.

- ❑ The observations should be in rows and the features in columns.

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.88	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.88	"USA"

Remove Duplicates

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.88	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.88	"USA"

- Lis appears twice.

- ❑ Clean data should not have duplicates.

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.88	"USA"
Hadrien		1.80	"FR"

Preparation, Exploration and Visualization

Use Unique ID

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.88	"USA"
Hadrien		1.80	"FR"

- The second Lis entry was removed, but it could be a valid entry.

❑ Each observation must have a unique ID.

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.88	"USA"
2	Hadrien		1.80	"FR"

Ensure Consistency

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.88	"USA"
2	Hadrien		1.80	"FR"

- Size seems to be in different measurement units.

❑ Measurements should use consistent units.

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Preparation, Exploration and Visualization

Ensure Homogeneity

ID	Name	Age	Size	Country
0	Sara	“27”	1.77	“Belgium”
1	Lis	“30”	1.70	“USA”
2	Hadrien		1.80	“FR”

- Two of the countries are abbreviated, one is not.

❑ Entries must be homogenous.

ID	Name	Age	Size	Country
0	Sara	“27”	1.77	“BE”
1	Lis	“30”	1.70	“USA”
2	Hadrien		1.80	“FR”

Correct Data Types

ID	Name	Age	Size	Country
0	Sara	“27”	1.77	“BE”
1	Lis	“30”	1.70	“USA”
2	Hadrien		1.80	“FR”

- “Age” is encoded as text.

❑ Correct data types must be used.

ID	Name	Age	Size	Country
0	Sara	27	1.77	“BE”
1	Lis	30	1.70	“USA”
2	Hadrien		1.80	“FR”

Preparation, Exploration and Visualization

Correct Missing Values

ID	Name	Age	Size	Country
0	Sara	27	1.77	“BE”
1	Lis	30	1.70	“USA”
2	Hadrien		1.80	“FR”

- Hadrien’s Age is missing.

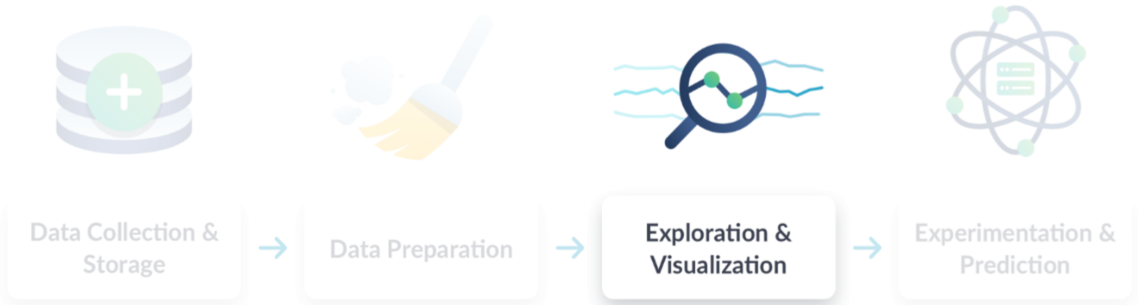
ID	Name	Age	Size	Country
0	Sara	27	1.77	“BE”
1	Lis	30	1.70	“USA”
2	Hadrien	28	1.80	“FR”

Reasons for Missing Values
- data entry
- error
- valid missing value

Solutions
- impute
- drop
- keep

	Sara	Lis	Hadrien	Lis
Age	“27”	“30”		“30”
Size	1.77	5.88	1.80	5.58
Country	“Belgium”	“USA”	“FR”	“USA”

Preparation, Exploration and Visualization



EXPLORATORY DATA ANALYSIS

Exploring the data → Formulating hypotheses → Assessing characteristics → Visualizing

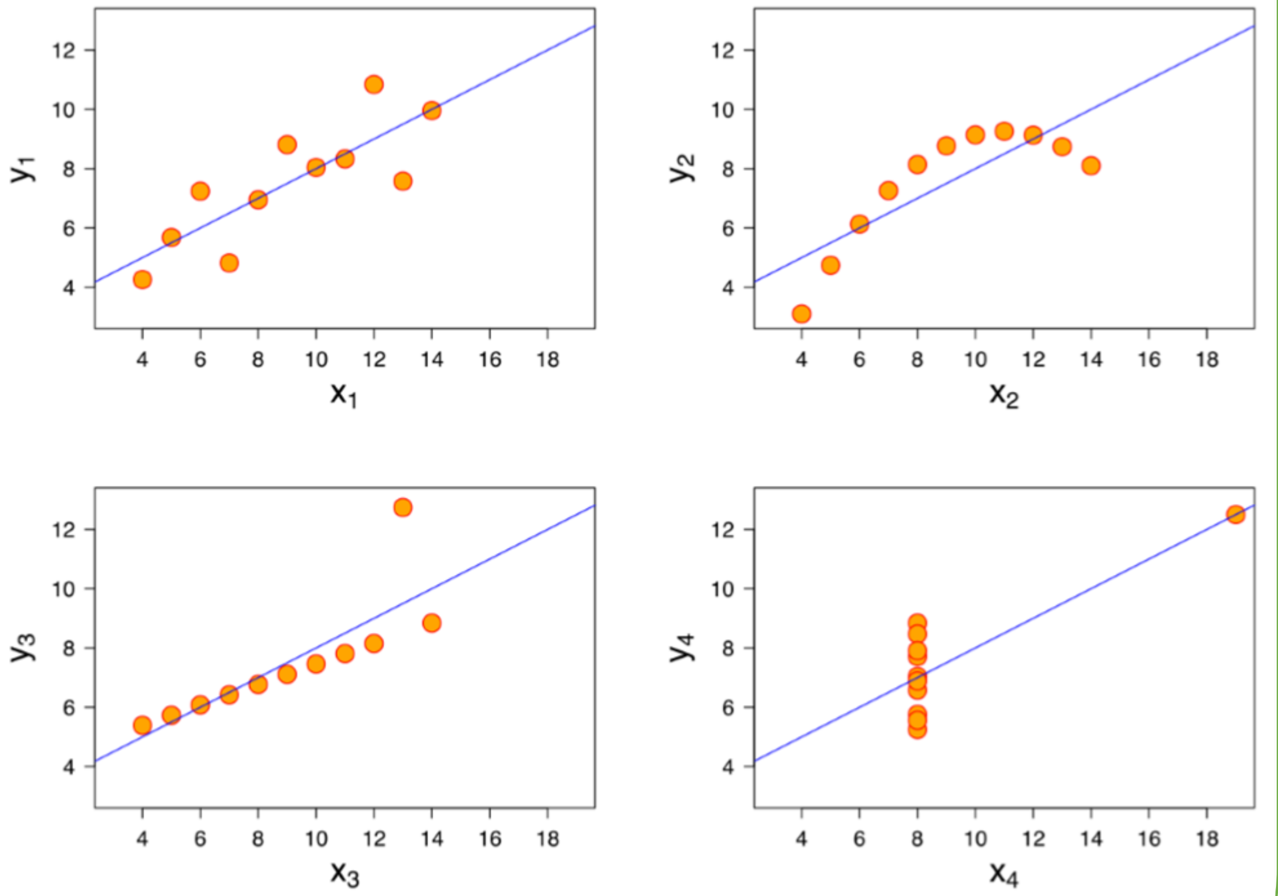
Preparation, Exploration and Visualization

Anscombe's Quartet

Dataset 1 x	Dataset 1 y	Dataset 2 x	Dataset 2 y	Dataset 3 x	Dataset 3 y	Dataset 4 x	Dataset 4 y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

$N = 11$ $\quad mean\,(x) = 9$ $\quad mean\,(y) = 7.5$

$r = 0.82$ $\quad sd\,(x) = 3.32$ $\quad sd\,(y) = 2.03$



Preparation, Exploration and Visualization

Illustrative Example. SpaceX Launches Dataset

1. Data Preview

Flight	Date	Time (UTC)	Booster Version	Launch Site	Payload
1	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Uni
2	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats
3	2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2+
4	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
5	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

Payload Mass (kg)	Orbit	Customer	Mission Outcome	Landing Outcome
NaN	LEO	SpaceX	Success	Failure (parachute)
NaN	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
525	LEO (ISS)	NASA (COTS)	Success	No attempt
500	LEO (ISS)	NASA (CRS)	Success	No attempt
677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Flight Number (number)
- Date (datetime)
- Time (datetime)
- Booster Version (text)
- Launch Site (text)
- Payload (text)
- Orbit (text)
- Customer (text)
- Mission Outcome (text)
- Landing Outcome (text)

Preparation, Exploration and Visualization

Illustrative Example. SpaceX Launches Dataset

2. Descriptive Statistics

	Flight	Date	Time (UTC)	Booster Version	Launch Site	Payload
count	55	55	55	55	55	55
unique	55	55	53	51	4	55
top	6	2018-03-30	4:45:00	F9 v1.1	CCAFS LC-40	SES-9
freq	1	1	2	5	26	1

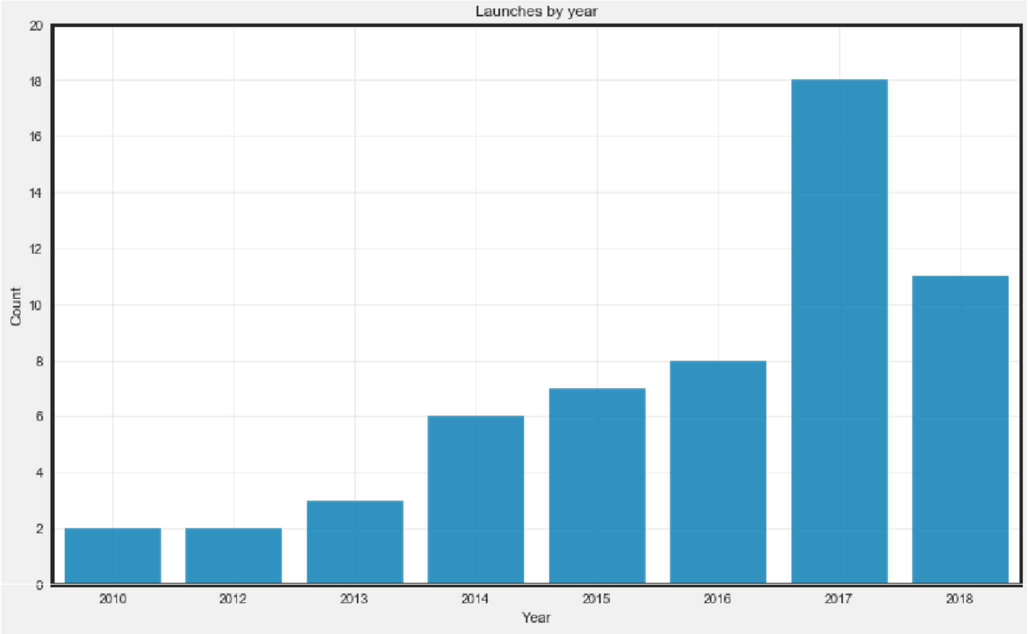
	Payload Mass (kg)	Orbit	Customer	Mission Outcome	Landing Outcome
count	53	55	55	55	55
unique	47	8	28	2	12
top	9,600	GTO	NASA (CRS)	Success	No attempt
freq	5	22	14	54	18

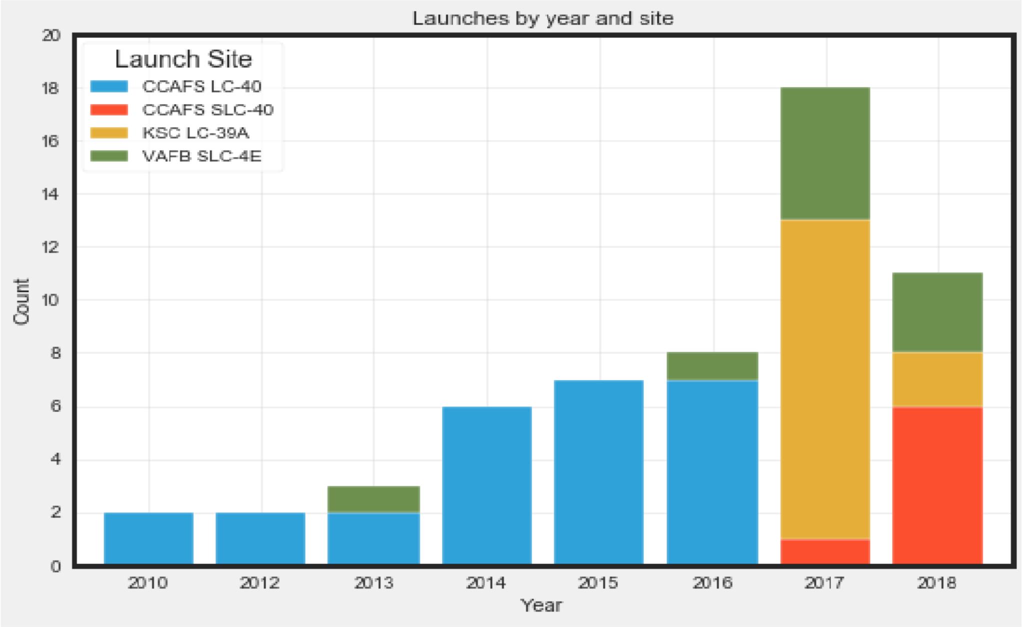
- 55 launches
- 2 missing values in the Payload Mass column
- 1 failed mission

Preparation, Exploration and Visualization

<u>Illustrative Example. SpaceX Launches Dataset</u>

3. Visualization





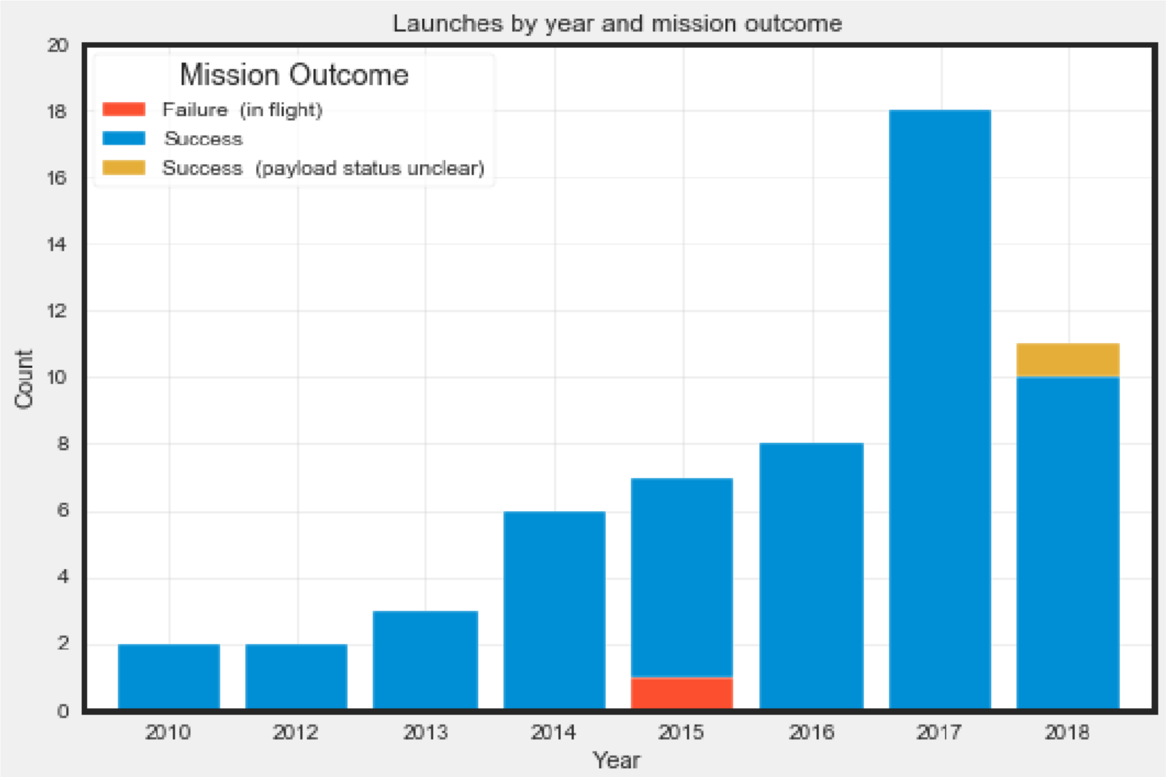
- No launch in 2011
- Gradual increase in launches before doubling in 2017

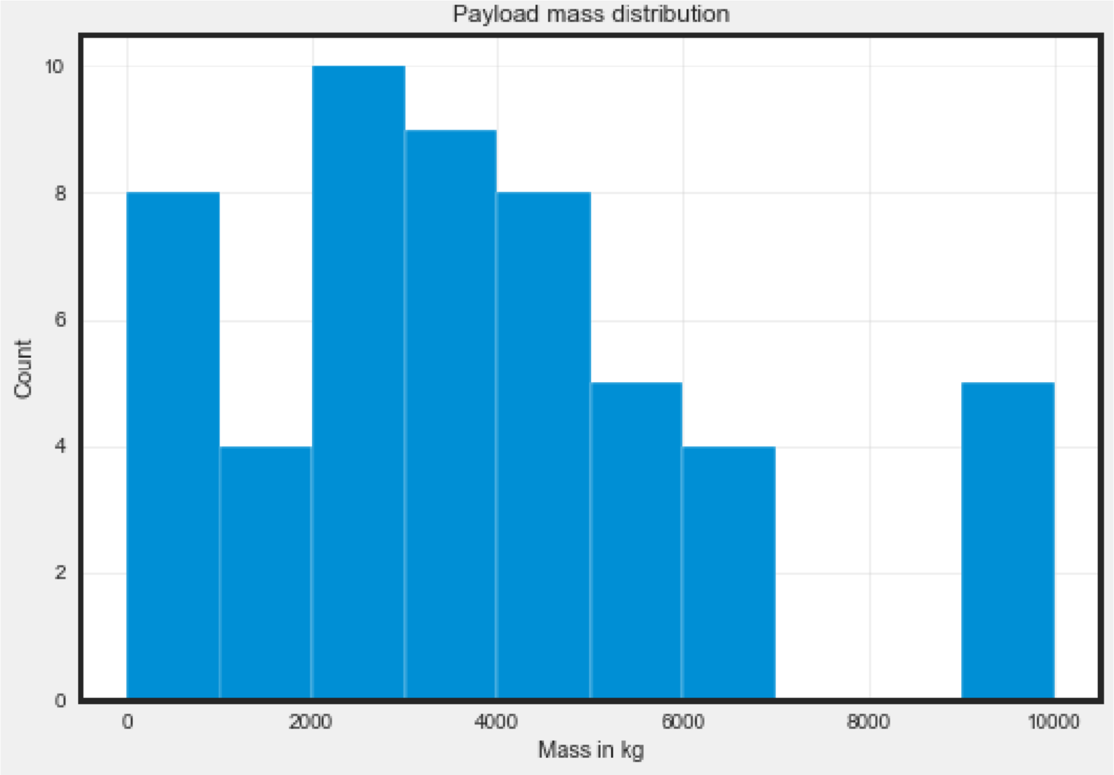
- Prior to 2017, launches originate from Cape Carnaveral Air Force Station (CCAFS)
- In 2017, most rockets launched from Kennedy Space Center (KSC)

Preparation, Exploration and Visualization

Illustrative Example. SpaceX Launches Dataset

3. Visualization

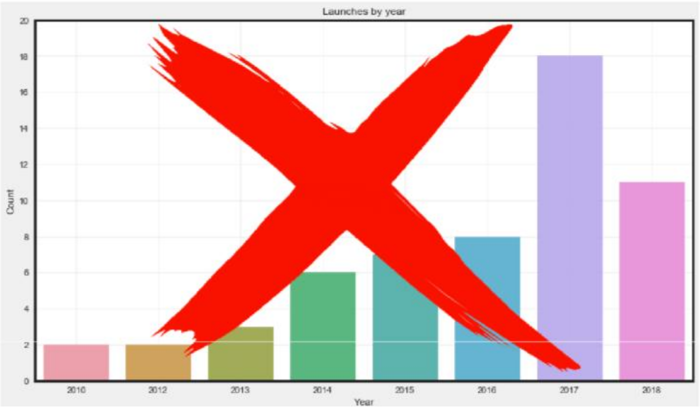


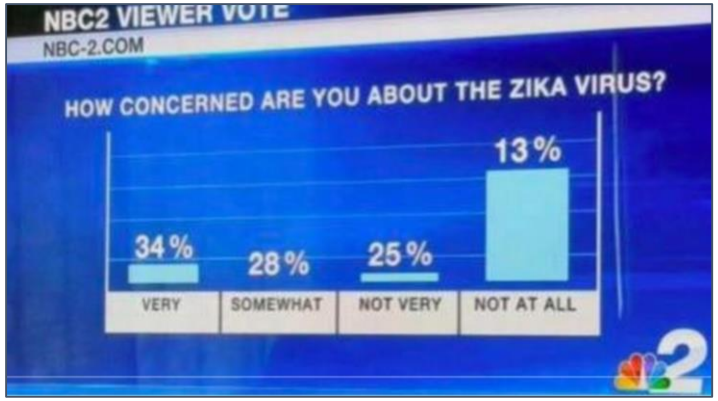


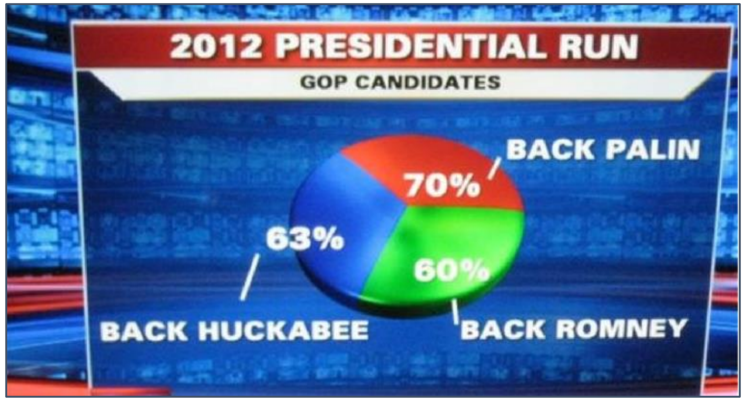
Preparation, Exploration and Visualization

Notes on Plot Preparation

- Use color purposefully
- Use color palettes that are distinguishable, even by color-blinded people
- Use readable fonts (sans serif)
- Use titles, axes labels and legends



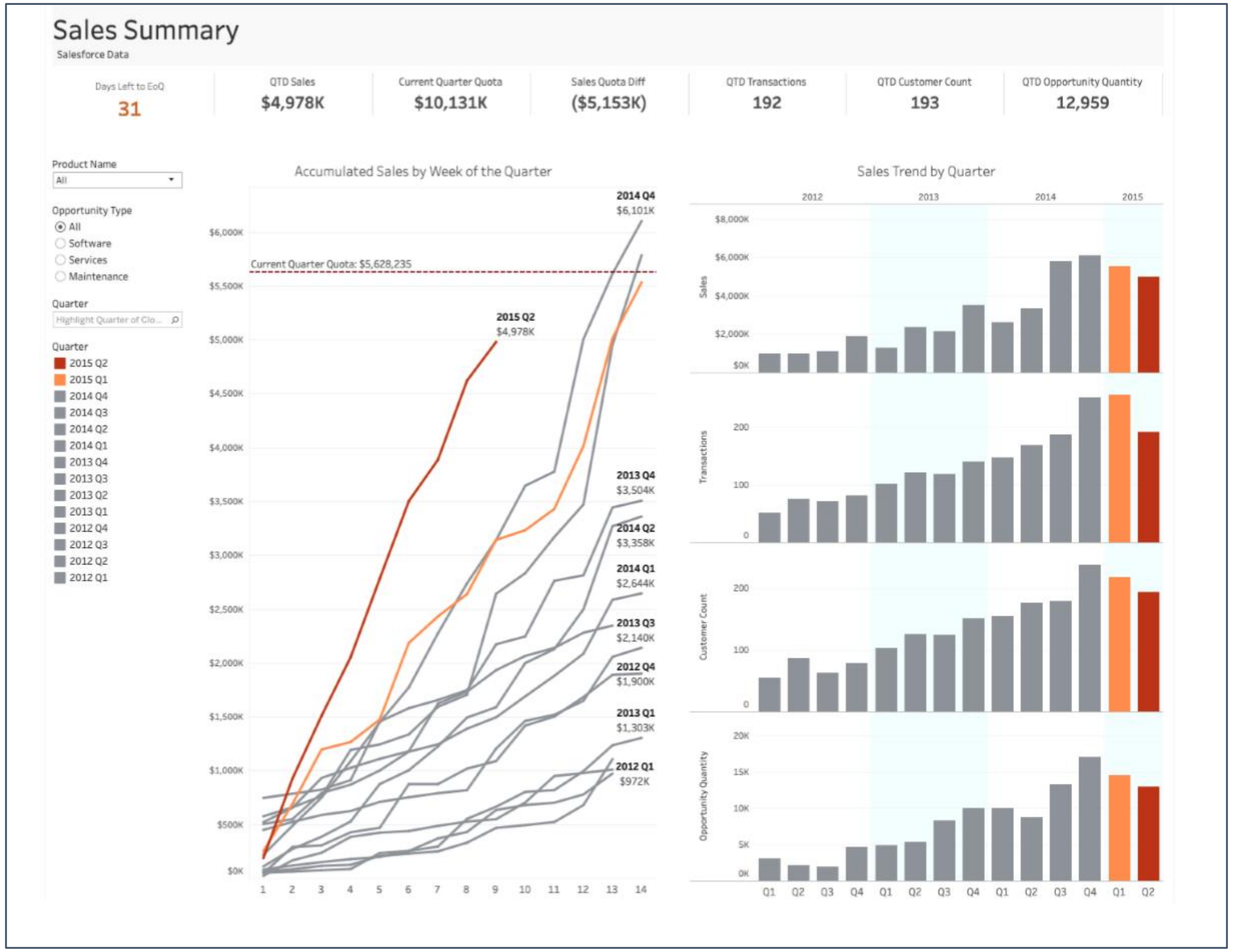




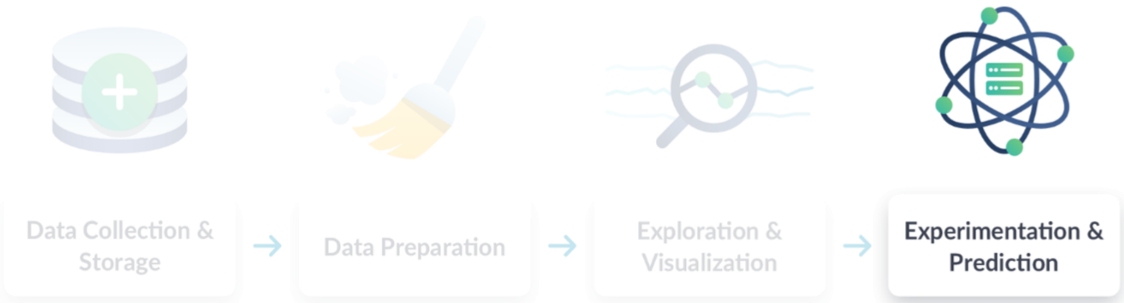
Preparation, Exploration and Visualization

DASHBOARDS

- group all the relevant information in one place to make it easier to gather insights and act on them



Experimentation and Prediction

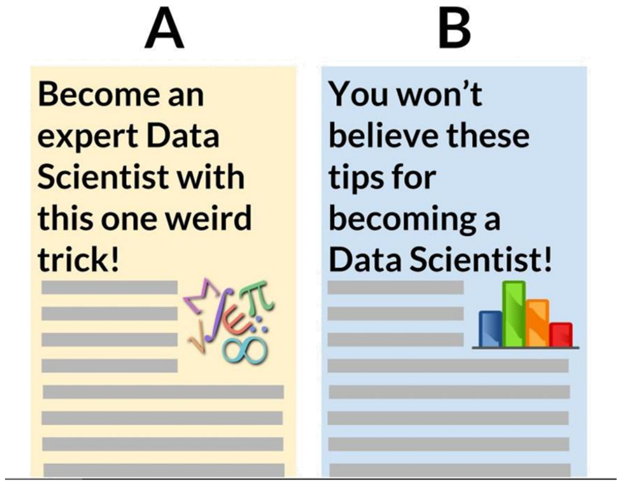


Experimental Procedure

1. State the question / problem
2. Formulate hypothesis
3. Collect data
4. Test the hypothesis with a statistical test
5. Interpret results

Experimentation and Prediction

<u>Case Study: Which is the better blog title?</u>



State the question / problem	Does blog title A or blog title B result in more clicks?
Formulate hypothesis	Blog title A and title B result in the same amount of clicks.
Collect data	50% of users will see blog title A, 50% title B. Track click-through rate until sample size is reached.
Test hypothesis	Is the difference in titles' clilck-through rates significant?
Interpret results	Choose title. Or design another experiment.

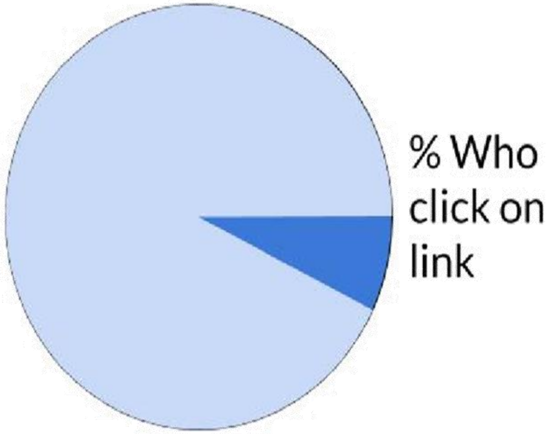
Experimentation and Prediction

A/B TESTING

- also called Champion / Challenger Testing
- used to make a choice between two options

Pick metric → Calculate sample size → Run the experiment → Check for significance

Pick metric → Click-through rate

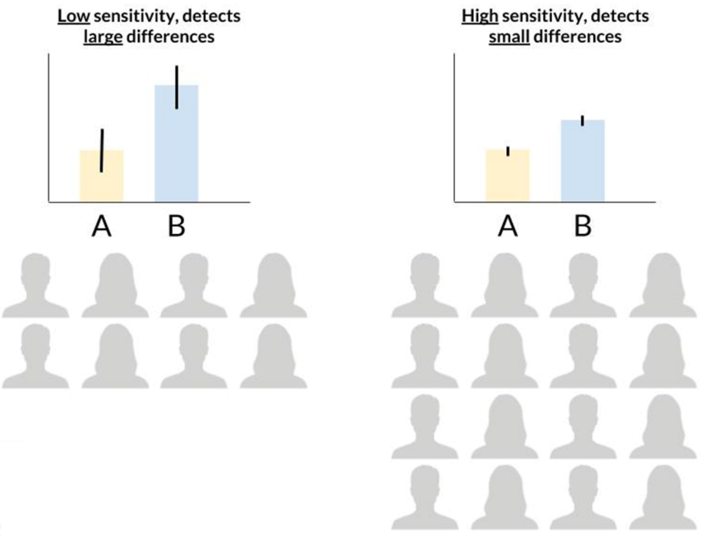


Experimentation and Prediction

A/B TESTING

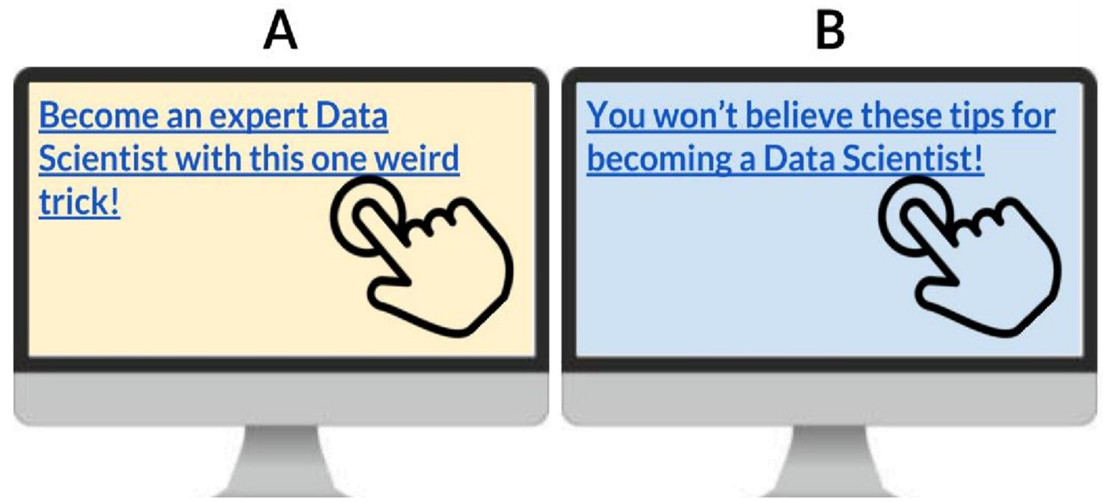
Calculate sample size

Larger sample sizes allow us to detect smaller changes.



Run the experiment

The experiment is run until the sample size is reached.

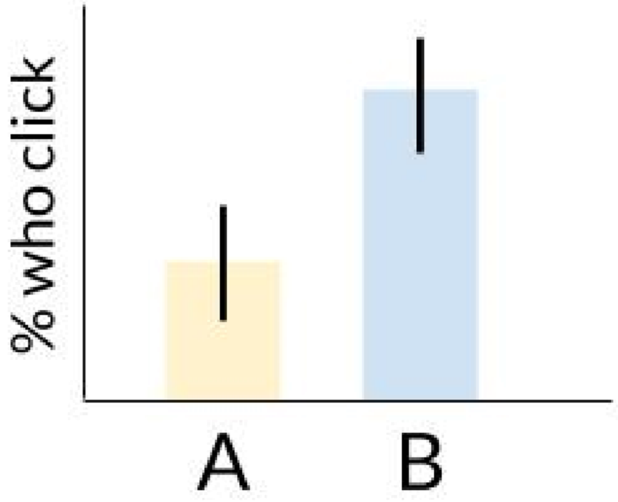


Experimentation and Prediction

A/B TESTING

Check for significance →

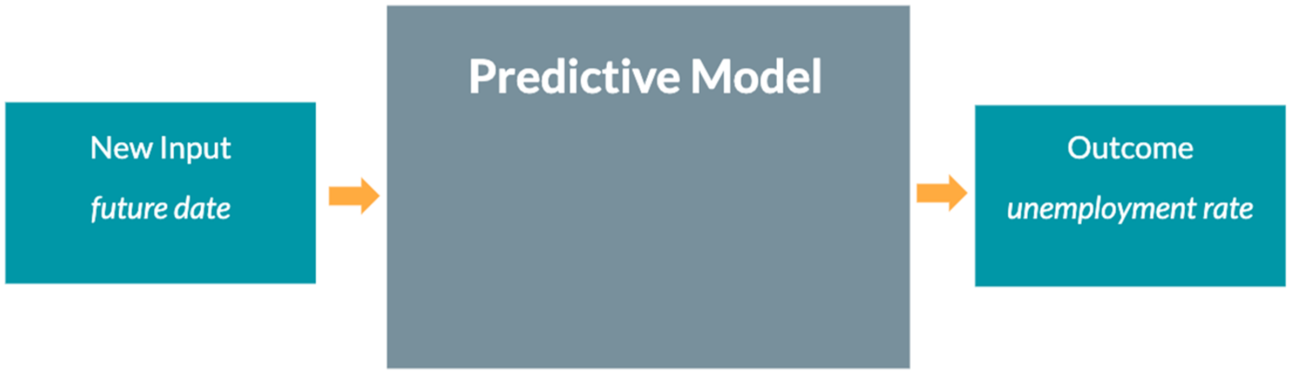
If the difference is significant, we can be reasonably sure that the difference is not due to random chance, but to an actual difference in preference.



Experimentation and Prediction

PREDICTIVE MODELING

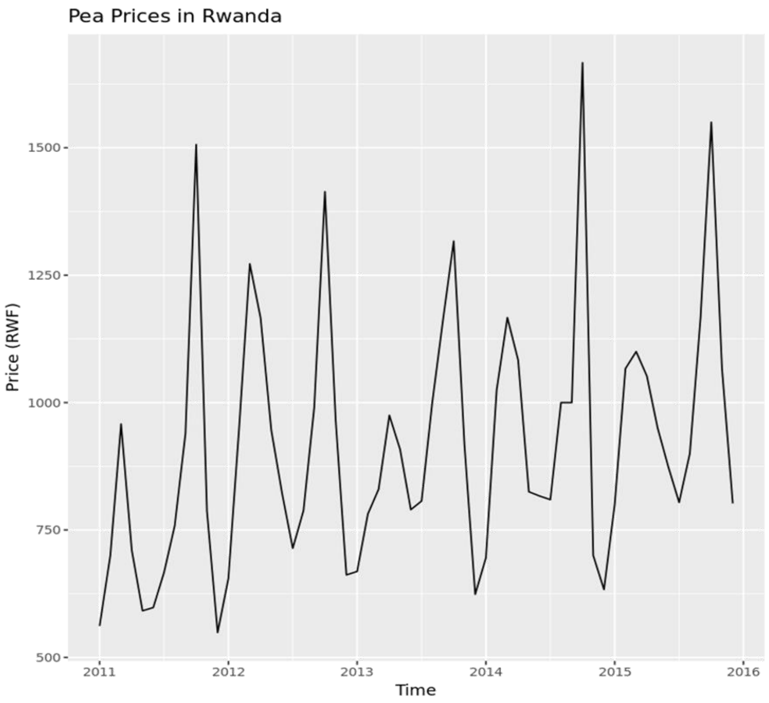
- by modeling a process, we can enter new inputs and see what outcome it outputs.



Experimentation and Prediction

PREDICTIVE MODELING

Case Study: Forecasting Time Series

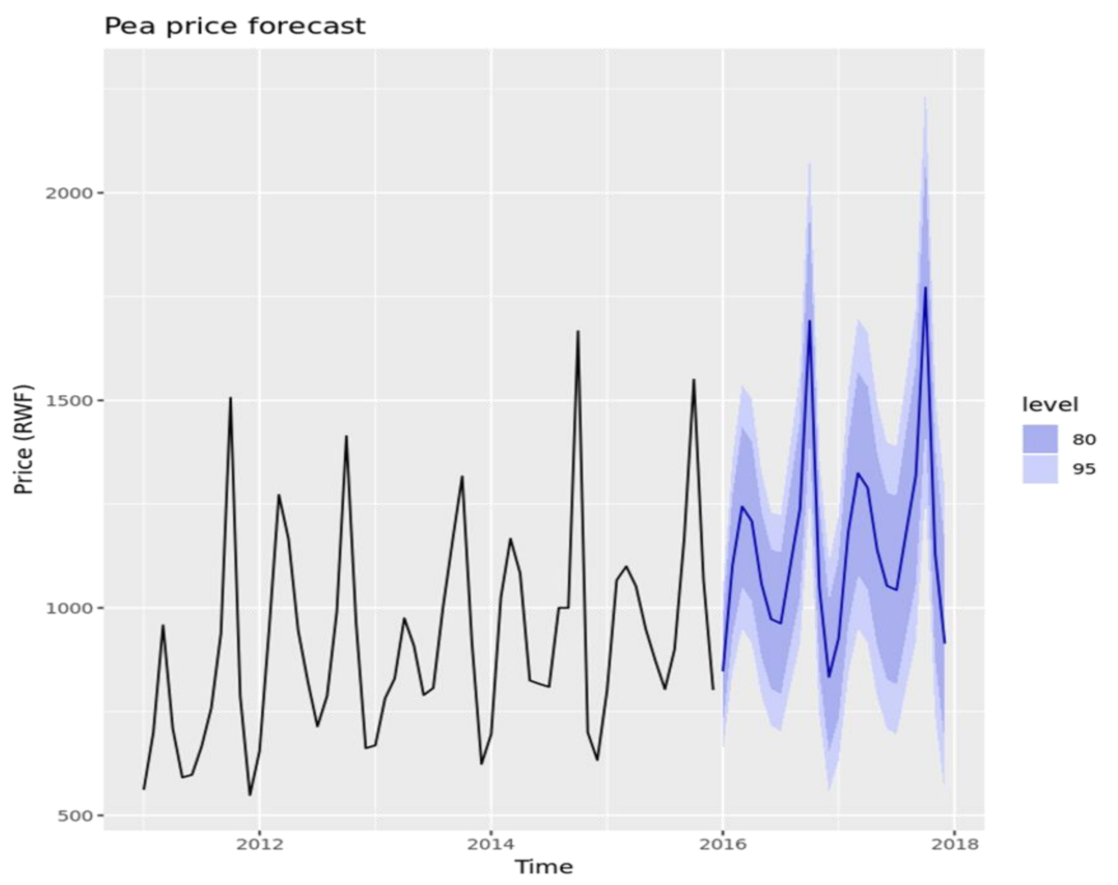


- Historical data for price of peas in Rwandan Francs from 2011 to 2016
- Seasonality: prices are lowest around December and January and peak around August; some years show a second peak around April
- General increase in pea prices annually

Experimentation and Prediction

PREDICTIVE MODELING

Case Study: Forecasting Time Series



- Confidence Interval: Model is X% sure that the true value will fall in this area

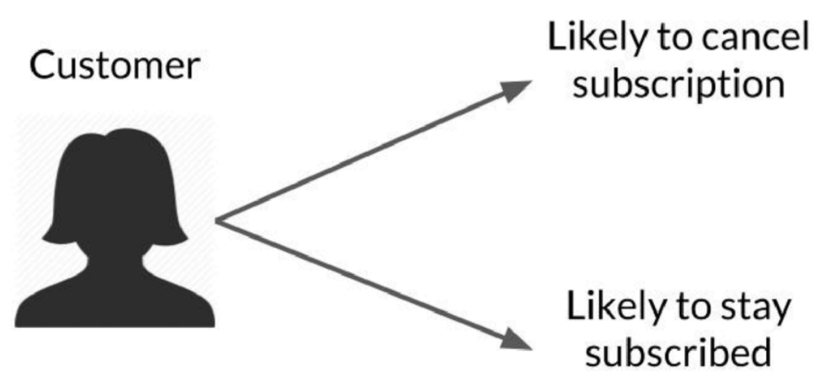
- The blue line depicts the forecast.
- The seasonality remains and it anticipates a continued increase in pea prices, ssen by the higher peaks and lows.

Experimentation and Prediction

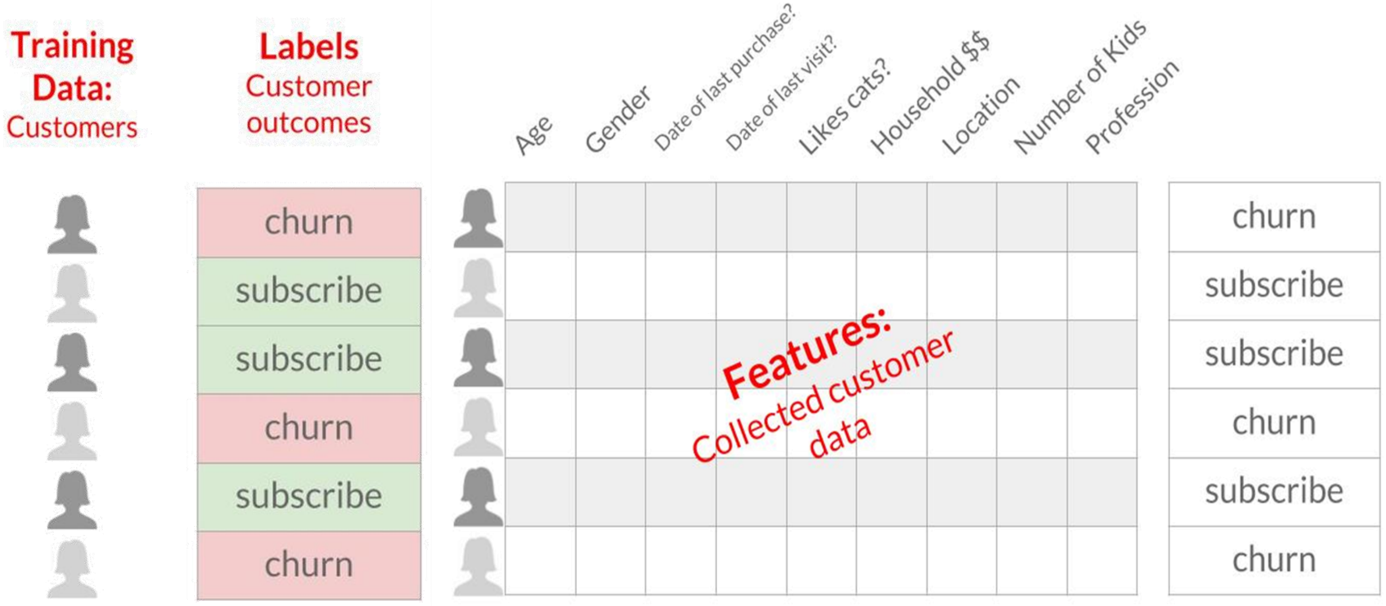
SUPERVISED MACHINE LEARNING

- use existing structured data to make prediction/s
- used for recommendation systems, diagnosing biomedical images, recognizing hand-written digits, predicting customer churn.

<u>Case Study: Predicting Customer Churn</u>



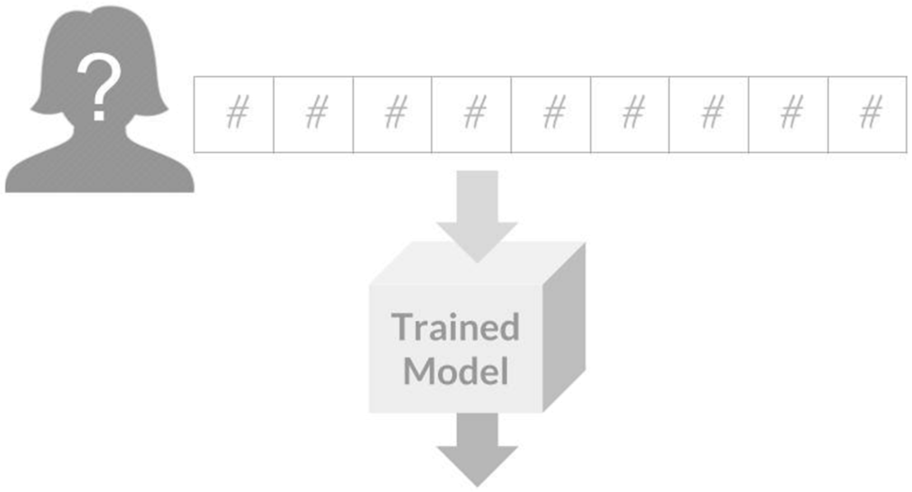
1. Model Preparation



Experimentation and Prediction

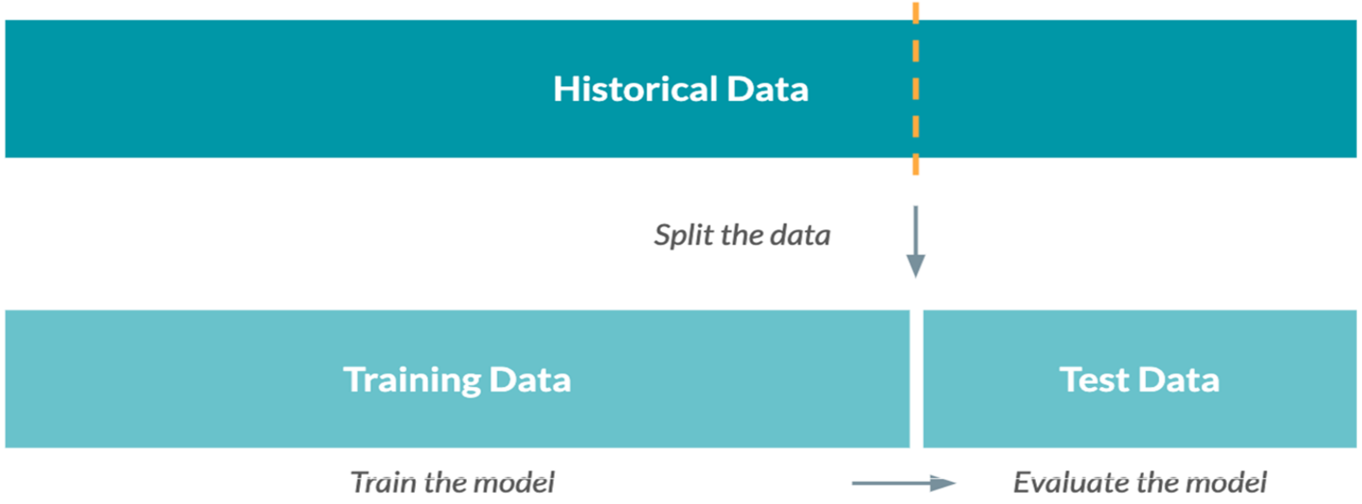
SUPERVISED MACHINE LEARNING

2. Model Use



Prediction: Subscribe

3. Model Evaluation



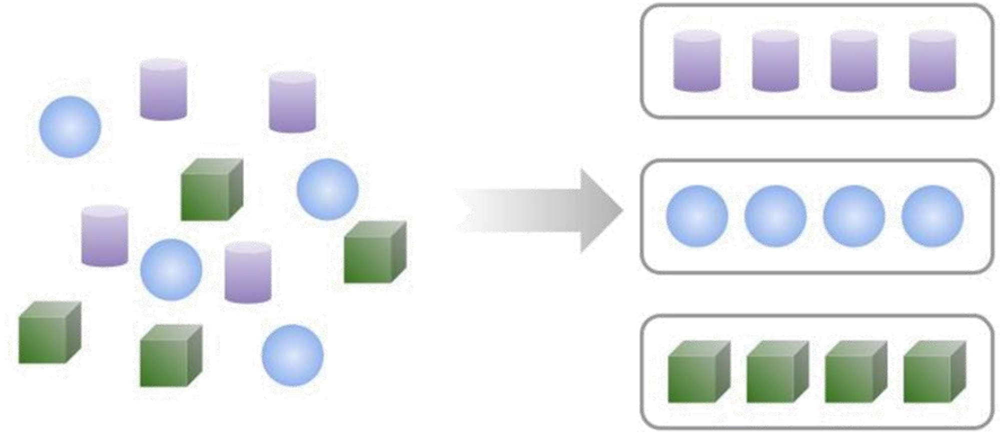
Possible Labels	True Labels	Model Prediction	Model Accuracy
Customer remains	970	1000	# of correct predictions / # of predictions = 970/1000 = 97%
Customer churns	30	0	

Experimentation and Prediction

CLUSTERING

- used to divide data into clusters

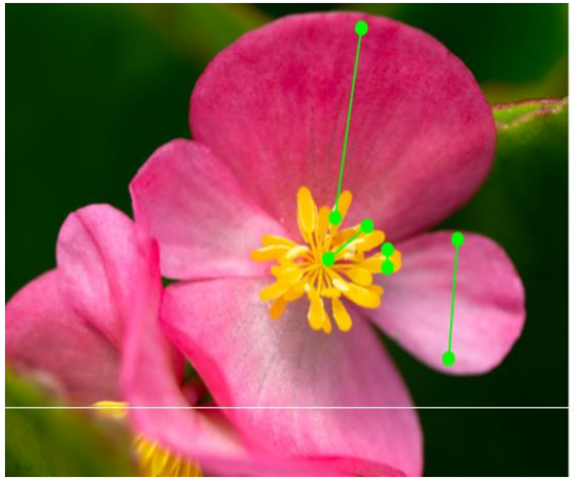
Select features → Select number of clusters → Use clusters to solve problems

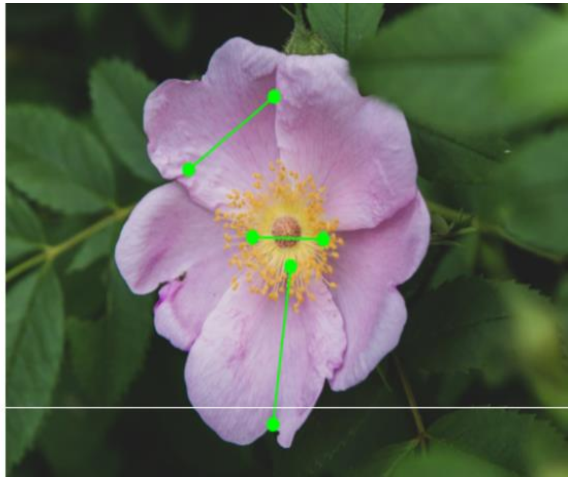


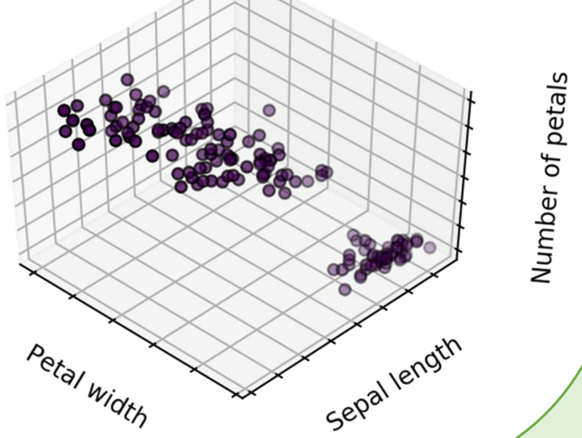
Case Study: Discovering New Species

Select features

- Flower colors
- Petal length and width
- Sepal length and width
- Number of petals







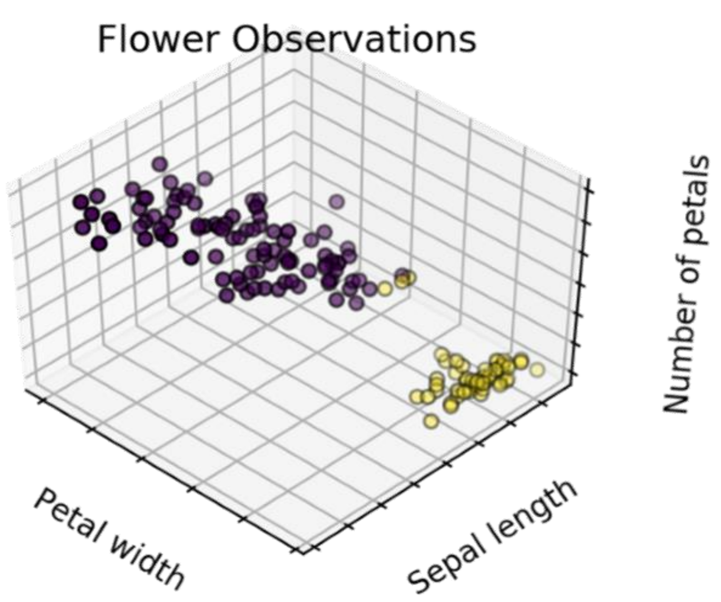
Experimentation and Prediction

CLUSTERING

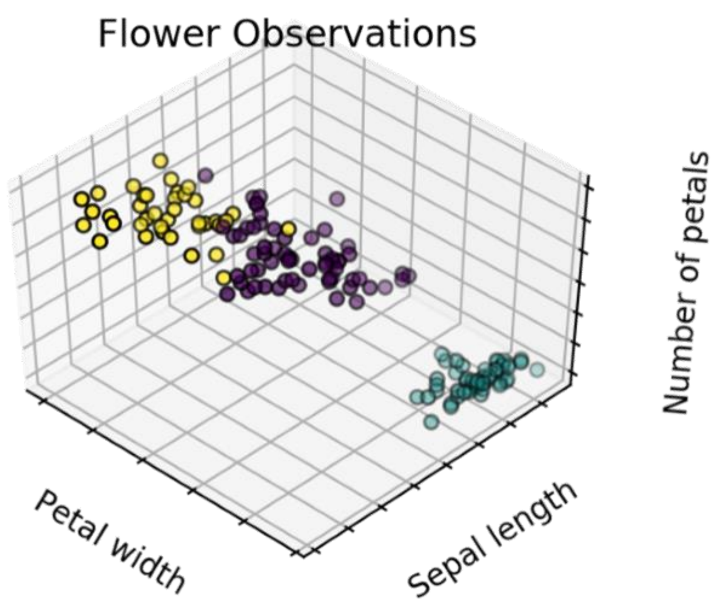
<u>Case Study: Discovering New Species</u>

Select number of clusters → The user eventually decides the final number of clusters; domain knowledge is important in deciding this

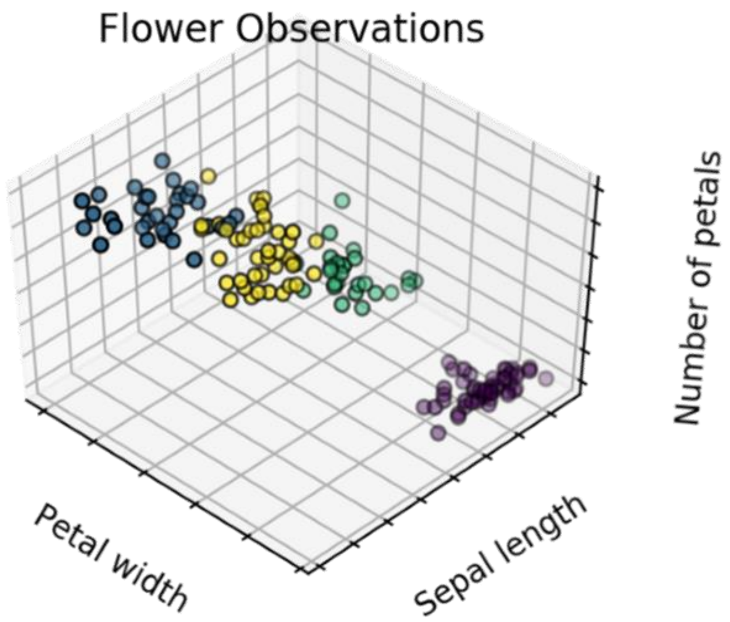
Two clusters



Three clusters



Four clusters



Case Study: Global Innovation Network and Analysis (GINA)

BACKGROUND

GINA: group of senior technologists located in centers of excellence (COEs) around the world

Charter: to engage employees across global COEs to drive innovation, research and university partnerships

New Director's Objectives:

- to improve these activities and provide a mechanism to track and analyze the related information
- to create more robust mechanisms for capturing the results of its informal conversations with other thought leaders within EMC, in academia, or in other organizations

Plan:

- provide a means to share ideas globally and increase knowledge sharing among GINA members who may be separated geographically
- create a data repository containing both structured and unstructured data to accomplish the goals

Goals

- Store formal and informal data
- Track research from global technologists
- Mine the data for patterns and insights to improve the team's operations and strategy

Case Study: Global Innovation Network and Analysis (GINA)

DISCOVERY

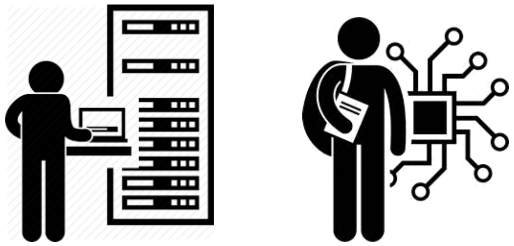
Business User
Project Sponsor
Project Manager

Vice President from the Office of the CTO

Business Intelligence Analyst

Representatives from IT

Data Engineer
Database Administrator



Representatives from IT

Data Scientist

Distinguished Engineer

Case Study: Global Innovation Network and Analysis (GINA)

DISCOVERY

- Data Categories
  - Innovation Roadmap
    - 5 years of idea submissions from internal innovation contests
    - SD: idea counts, submission dates, inventor names
    - UD: textual description of the ideas
  - Minutes
    - notes representing innovation and research activities from around the world
    - SD: dates, names and geographic locations
    - UD: rich data about knowledge growth and transfer

Case Study: Global Innovation Network and Analysis (GINA)

DISCOVERY

	INITIAL HYPOTHESES
A	Descriptive analytics of what is currently happening can spark further creativity, collaboration and asset generation.
B	Predictive analytics can advise executive management of where it should be investing in the future.
1	Innovation activity in different geographic regions can be mapped to corporate strategic directions.
2	The length of time it takes to deliver ideas decreases when global knowledge transfer occurs as part of the idea delivery process.
3	Innovators who participate in global knowledge transfer deliver ideas more quickly than those who do not.
4	An idea submission can be analyzed and evaluated for the likelihood of receiving funding.
5	Knowledge discovery and growth for a particular topic can be measured and compared across geographic regions.
6	Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.
7	Strategic corporate themes can be mapped to geographic regions.
8	Frequent knowledge expansion and transfer events reduce the time it takes to generate a corporate asset from an idea.
9	Lineage maps can reveal when knowledge expansion and transfer did not (or has not) resulted in a corporate asset.
10	Emerging research topics can be classified and mapped to specific ideators, innovators, boundary spanners and assets.

Case Study: Global Innovation Network and Analysis (GINA)

DATA PREPARATION

Many of the names of the researchers and people interacting with the universities were misspelled or had leading and trailing spaces in the datastore.

MODEL PLANNING

- Social network analysis techniques
- Initiate a longitudinal study – to begin tracking data points over time regarding people developing new intellectual property

Considerations for the Parameters of the Longitudinal Study

- Identify the right milestones to achieve this goal.
- Trace how people move ideas from each milestone toward the goal.
- Trace ideas that die, and trace others that reach the goal. Compare the journeys of ideas that make it and those that do not.
- Compare the times and the outcomes using a few different methods (depending on how the data is collected and assembled).

Case Study: Global Innovation Network and Analysis (GINA)

MODEL BUILDING

- Natural Language Processing (NLP) for the textual descriptions in the Innovation Roadmap ideas
- Social network analysis using R and RStudio

Social graph showing submitters and finalists

Case Study: Global Innovation Network and Analysis (GINA)

MODEL BUILDING

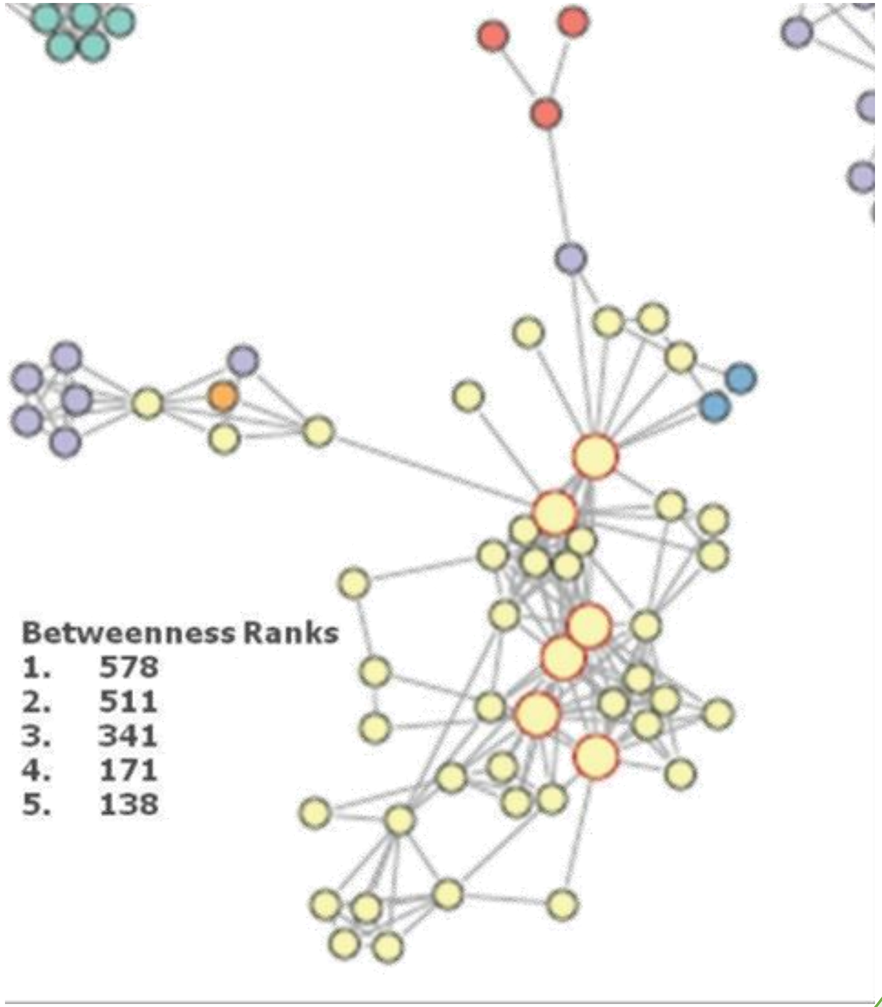
Each color represents an innovator from a different country. The large dots with red circles around them represent hubs (person with high connectivity and a high "betweenness" score.

One person has an unusually high score. A query on him yielded information on his attendance to conferences (at different locations) and interactions with other scientists.

The finding suggests that at least part of the initial hypothesis is correct: the data can identify innovators who span different geographies and business units.

Software / Database Used:

- Tableau – for data visualization and exploration
- Pivotal Greenplum Database – main data repository



Case Study: Global Innovation Network and Analysis (GINA)

RESULTS

The project was considered successful in identifying boundary spanners and hidden innovators. As a result, the CTO office launched longitudinal studies to begin data collection efforts and track innovation results over longer periods of time.

Applying social network analysis enabled the team to find a pocket of people within EMC who were making disproportionately strong contributions. These findings were shared internally through presentations and conferences and promoted through social media and blogs.

OPERATIONALIZATION

Key Findings:

- The CTO office and GINA need more data in the future, including a marketing initiative to convince people to inform the global community on their innovation / research activities.
- Some of the data is sensitive, and the team needs to consider security and privacy related to the data, such as who can run the models and see the results.
- In addition to running models, a parallel initiative needs to be created to improve basic Business Intelligence activities, such as dashboards, reporting, and queries on research activities worldwide.
- A mechanism is needed to continually reevaluate the model after deployment.

Case Study: Global Innovation Network and Analysis (GINA)

SUMMARY

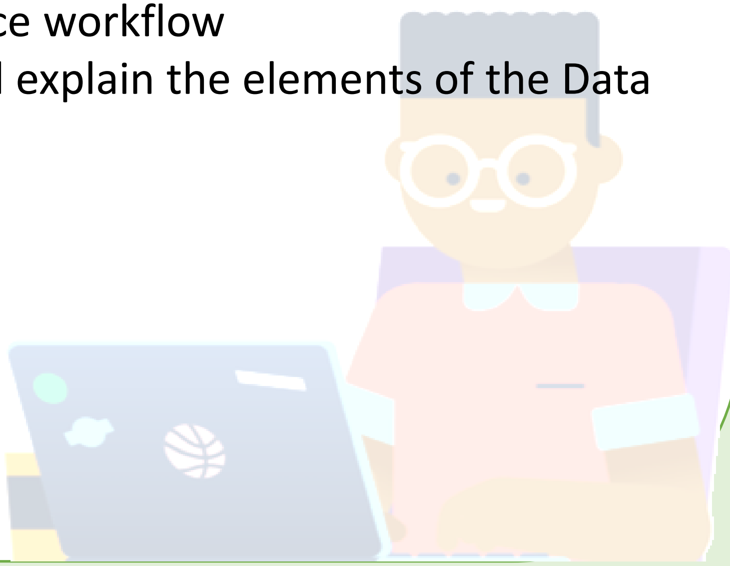
Component	Result/s
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypothesis	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis
Result and Key Findings	1. Identified hidden, high-value innovators and found ways to share their knowledge 2. Informed investment decisions in university research projects 3. Created tools to help submitters improve ideas with idea recommender systems

Outline

Module 1.1: INTRODUCTION TO DATA AND DATA SCIENCE
Data and the Data Ecosystem
Data Science and its Applications
Data Science Roles and Tasks
The Data Scientist
The Data Science Workflow
Data Collection and Storage
Preparation, Exploration and Visualization
Experimentation and Prediction
Case Study: Global Innovation Network and Analysis

Learning Outcomes

1. Define data, its properties, importance and capabilities
2. Explain the drivers of data and the current data ecosystem
3. Define Data Science and differentiate its applications
4. Differentiate the Data Science roles and enumerate the tools needed by each role
5. Explain the skills and characeristics that a Data Scientist must possess
6. Explain the phases of the Analytics lifecycle and relate these to the Data Science workflow
7. Identify, enumerate and explain the elements of the Data Science workflow



Introduction to Data and Data Science

Engr. Elisa G. Eleazar

School of Chemical, Biological, and Materials Engineering and Sciences